UNIVERSITY OF OSLO
Department of Informatics

Evaluation of gene prediction methods for prokaryotes

Master's thesis

Stian Engebretsen

15th August 2012

Abstract

There are currently over 2,000 completed prokaryotic sequences available. In these DNA sequences the genes lay hidden. The genes codes for proteins that forms the foundation for all organisms. Several programs that try to locate these genes have been developed. However, little has been done to evaluate these programs on a large scale. The goal of this thesis is to evaluate these programs closer, and develop guidelines for when to use the different programs.

Contents

List of figures

List of tables

Preface

First, I want to thank my supervisours Karin Lagesen and Anja Bråthen Kristoffersen for their guidance. Secondly, I want to thank the people at the research group for Biomedical Informatics (BMI) at the Department of Informatics. I also want to thank the people at the Centre of Ecological and Evolutionary Synthesis (CEES) at the Department of Biology. Lastly, I want to thank my family and friends.

Stian Engebretsen

Chapter 1

Introduction

SINCE THE GENOME of bacteriophage φX174 was sequenced by Sanger, Nicklen and Coulson in 1977, the speed that new genomes get sequenced has increased dramatically.

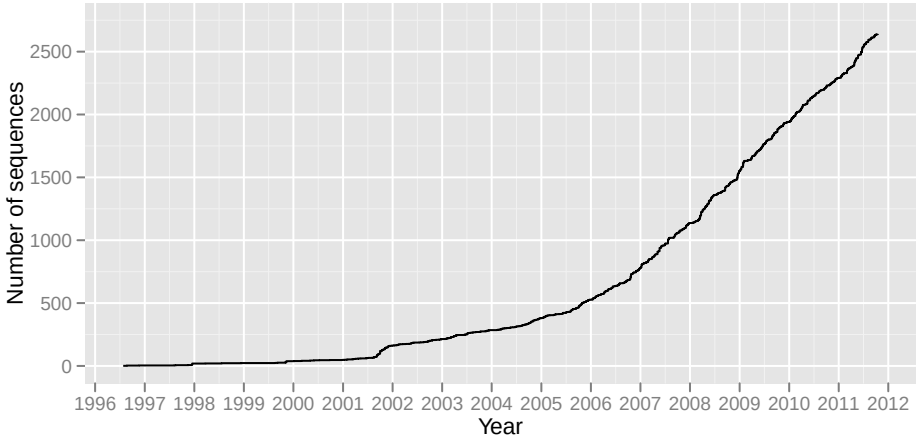
The first whole genome for a living organism — the bacterium *Haemophilus influenzae* — where sequenced by Fleischmann et al. in 1995. The first 100 finished sequences where sequenced by October 2001 (see figure 1.1), and already by August 2007 the first 1,000 finished sequences had been sequenced, and the number is still rapidly raising thanks to next-generation sequencing methods.

In order to deal with the high biological data throughput, the need for automatic annotation methods for newly sequenced genomes becomes apparent. This is especially needed for prokaryotic sequences given the rapidly rising number of prokaryotic genomes, due to their relative simplicity to sequence compared to the sequencing of genomes for eukaryotic species.

The first coding gene prediction programs appeared before even the first whole bacterial genome became available in 1995, and since then the number of available gene prediction methods has steadily increased as our understanding of the gene mechanics has improved.

This raises the question: *Which of these programs gives us the best results?* This master thesis attempts to shed some light on this question by comparing the five prokaryotic coding gene prediction programs GeneMark.hmm, GeneMarkS, Glimmer, MED and Prodigal.

Figure 1.1: Graph showing the number of finished sequences in NCBI Genome plotted against the date they were first added to NCBI Genome.
Based on data from the file
`ftp://ftp.ncbi.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt`



1.1 Research questions

- Are there any general differences in the performance of the different programs?
- Are there any specific situations were one (or more) program(s) are more suitable than others?
- Are there specific types of coding genes or other relevant features that the programs can't handle?

Chapter 2

Background theory

THIS CHAPTER CONTAINS background theory necessary for this thesis. First, the concepts of prokaryotic organisms and genes are explained. Then, some background theory on the prediction methods commonly used to predict genes is explained.

2.1 Prokaryotic genes

This section is based on the book *Brock Biology of Microorganisms* by Madigan et al. (2011).

2.1.1 Prokaryotic cells

The basic unit of life is the *cell*. A cell is separated from other cells by a *membrane*.

The cell is an open system where nutrients from the environment are taken up. The nutrients are then transformed into useful products, which is used by the cell, and waste products, which are released out to the environment. This process is know as *metabolism*.

The purpose of metabolism is *growth*. Growth is the term used when one cell divides and forms two cells.

In order to better adapt to the environment the cells live in, the new cells will also be modified in order to increase their *fitness*. This is know as evolution, and allows the cell to gain new biological properties.

Looking at the internal cell structures, two patterns emerge. One pattern is cells with a nucleus, the *eukaryotes*, while the other pattern is cells that lack a nucleus, the *prokaryotes*.

The prokaryotes are divided into the *Archaea* and the *Bacteria*. While they are both prokaryotes, they are *evolutionary distinct*.

Long ago, the last universal common ancestor divided into the *Bacteria* and a second common ancestor linage. Later, this second linage divided into *Archaea*, which is prokaryotic, and the *Eukarya*, which is not.

In other words, the *Archaea* is closer related to the *Eukarya* than the *Bacteria*, even though both *Bacteria* and *Archaea* are prokaryotes.

Table 2.1: An example showing the levels of taxonomy used in this thesis for *Deinococcus radiodurans*

Taxonomic level	Example value
Domain	*Bacteria*
Phylum	*Deinococcus-Thermus*
Class	*Deinococci*
Order	*Deinococcales*
Family	*Deinococcaceae*
Genus	*Deinococcus*
Species	*Deinococcus radiodurans*

2.1.2 Taxonomy

Taxonomy is the placing of organism into groups based on several defined characteristics of the organism. These groups consists of both high level ones such as *domain*, and low level one, such as *species*. In other words, in consists of several levels which is based on gradually more specific characteristics of the organism in question.

See table 2.1 for an example for the bacterium *Deinococcus radiodurans*.

2.1.3 The central dogma

The cells store the *genetic information* using deoxyribonucleic acid (DNA) in what is known as the *genome*.

The genome, on one hand, is *replicated* to form a copy of the genome. This copy is given to the new cell that is formed under growth. This is to ensure the new cell has its own copy of the genome, so that the new cell also can grow, and thus allow for the growth of cells until some factor limits their growth, such as the lack of nutrients.

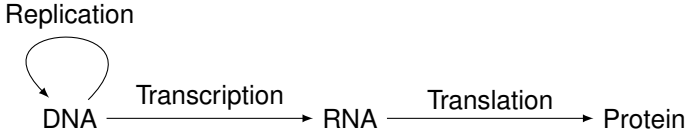
Certain regions of the genome is *transcribed* into ribonucleic acid (RNA), and these regions is said to code for genes.

The transcription is performed by RNA *polymerase*, which uses the DNA as a template for producing the RNA, in the case of coding genes, messenger RNA (mRNA).

The RNA polymerase attaches to the DNA strand in a region known as the *promoter* region for the RNA polymerase. This promoter region is located upstream of the sequence of the coding gene.

Once the RNA polymerase has started transcribing, it will continue to produce RNA until a stem-loop forms in the RNA. When this stem-

Figure 2.1: Schematic figure of the central dogma of molecular biology



loop forms, the RNA polymerase will stall, and as a result, transcription terminates.

The next step is *translation* which is the process of translating mRNA into protein. Translation starts by the *ribosome* attaching to the region of mRNA known as the Ribosome Binding Site (RBS). The RBS is located just upstream of the start codon. A codon is a triplet of nucleotides.

Once attached, the ribosome will start producing an amino acid chain starting from the *start codon*. For each codon, it will use transporter RNA (tRNA) to select the correct amino acid for the codon being translated. Which codon codes for which amino acid is specified in the *genetic code*, which is considered universal, except for some small variations for small parts of the genetic code. The translation will continue until the ribosome reaches the *stop codon*.

This process of DNA being transcribed into RNA, which is then translated into protein, is known as *the central dogma of molecular biology* (see figure 2.1).

2.1.4 Protein coding gene signals

The coding genes in a genome can be located by looking for regions that start with a *start codon* and end with a *stop codon*. Such regions are called Open Reading Frames (ORFs).

Simplified, the start codon is the location where the transcription begins. The transcription continues until the stop codon is reached. The region should not be interrupted by other stop codons. Start codons are not unusual inside genes.

The region from the start codon, until the stop codon, is the DNA sequence for the coding gene, and by using the genetic code, one can translate this sequence to the amino acid sequence of the protein.

The potential stop codon are easy to locate since the codons UGA, UAG, UAA are used only to code for stop[1].

The potential start codons, most commonly AUG, GUG, UUG[2], also codes for amino acids. This means the gene prediction programs need to determine if the potential start codon is the start codon, or a amino acid inside the gene. The result is that the start codon is harder to accurately predict than the stop codon.

However, not all ORFs are coding genes, and to improve the prediction results, one can look for additional features around the ORF. The most common feature to look for is the RBSs, which is located upstream of the start codon.

The RBS is the location where the ribosome binds to the mRNA during translation, and is located close to the start codon. If a coding gene has a RBS, it is likely to be a real coding gene.

Another way of increasing the accuracy of predictions is to compare the probability that the ORF is coding to the probability that the ORF is non-

[1] With some minor variation in special situations. See § 6.1.1.

[2] Other start codons are possible, but quite rare.

coding. If the probability that the ORF is coding is close to the probability that the ORF is non-coding, i.e., the bases in the ORF appear in a random fashion, the ORF is probably not a real coding gene.

To summarize, locating potential coding genes consists of looking for regions bounded by a start codon and stop codon with a RBS located upstream of the start codon and sometimes also coding bias.

2.2 Gene prediction methodologies

2.2.1 Markov models

A Markov model is a model that describes the process of moving from one state to another. The transcision from a state to another has a probability specified by the *transcision probabilities*. (Xiong 2006)

Interpolated Markov Model (IMM) When using high-order Markov model, a problem that might arise is the lack of sufficient $n$mers to train the model of order $n-1$. As a result, the performance of the high-order model might suffer on short sequences. This problem can be solved by using an IMM which is a variable-length Markov model. (Xiong 2006)

Hidden Markov Model (HMM) As with the Markov models, the HMM also models the transcision from one state to another using transcision probabilities. However, in HMMs there also exists some unobservable factors that influence the transcisions. The unobservable factors are modeled using *hidden states*. (Xiong 2006)

The *Viterbi algorithm* is used to find the most probable state path for HMMs.

2.3 Gene prediction programs

2.3.1 Introduction

The purpose of (coding) gene prediction programs is to correctly predict the protein coding genes (see § 2.1) in a given sequence. While the initial process of finding candidate genes is similar (see § 2.1.4), the methods for improving this base prediction differs between the different programs.

However, there exists some similarities between the programs when it comes to the refinement of the base predictions. Three of the programs — GeneMark.hmm (Lukashin and Borodovsky 1998), GeneMarkS (Besemer, Lomsadze and Borodovsky 2001) and Glimmer (Delcher et al. 2007) — uses a Markov model (see § 2.2.1) based architechture.

The two remaining programs — MED (Zhu et al. 2007) and Prodigal (Hyatt et al. 2010) — is based on different approches. MED uses an Entropy Density Profile (EDP) model, while Prodigal is based on dynamic programming.

There also exists common problems all the programs must handle for producing accurate results. Some of these problems include the ability to handle overlapping genes and (real) short genes. Another problem for ensuring accurate start site prediction include handling genes using different RBS usage than the Shine-Dalgarno (SD) motif (Shine and Dalgarno 1975). They must also handle accurate predictions on both AT-rich and GC-rich genomes (see § 5.2.3).

2.3.2 GeneMark.hmm

This section is based on Lukashin and Borodovsky (1998).

The gene prediction program GeneMark.hmm is uses an architecture based on HMM with duration and nine hidden states. For the purpose of gene predicting, GeneMark.hmm uses an modified version of the Viterbi algorithm. The algorithm is modified to handle variable duration.

GeneMark.hmm did not handle overlapping genes back in 1998, but a newer version supports overlaps of arbritatry lengths (see § 2.3.3).

A problem arising when using the Viterbi algorithm for gene prediction is that it has a tendency to predict overlapping genes shorter than they are. This is fixed in GeneMark.hmm by using a RBS post-processing step that picks an alternative gene start if it scores above a certain threshold. If no such start is found, the Viterbi prediction is used.

2.3.3 GeneMarkS

This section is based on Besemer, Lomsadze and Borodovsky (2001).

The gene prediction program GeneMarkS combines an improved version of GeneMark.hmm with heuristic Markov models and the Gibbs sampling method.

The improved version of GeneMark.hmm can handle gene overlaps of arbritatry length. In addition, the improved version includes RBS related functions in the Viterbi algorithm.

The RBS models used are based on Gibbs sampling.

GeneMarkS uses the heuristic Markov models to run GeneMark.hmm. It uses the prediction output from one step to build new models. It then produces new predictions, then iterates again. This process is repeated until the predictions between two steps are sufficently close.

2.3.4 Glimmer

This section is based on Delcher et al. (2007).

The gene prediction program Glimmer uses an IMM based architecture. This scores all ORFs in reverse. This provides a more accurate score near the start codon.

Improvement of the start prediction can be perform using the external program ELPH[3]. This program uses Gibbs sampling.

[3] `http://cbcb.umd.edu/software/ELPH`

It is also possible to combine Glimmer and ELPH in an iterative fashion. This is similar to what is done in GeneMarkS.

Glimmer uses dynamic programming to select ORFs. This reduces the number of false positives that appear due to overlapping ORFs.

2.3.5 MED

This section is based on Zhu et al. (2007).

The gene prediction program MED uses an Entropy Density Profile (EDP) based architecture. The MED algorithm is based on the assumption that the EDP-vector for coding and non-coding ORFs forms separate clusters in the EDP space. This means that an ORF can be determined to be coding or not, by finding which cluster the ORF is nearest.

The EDP clusters for coding and non-coding ORFs are thought to be universal, and thus this part of the algorithm requires no training.

MED also has an RBS improvement step based on Position Weight Matrix (PWM).

2.3.6 Prodigal

This section is based on Hyatt et al. (2010).

The gene prediction program Prodigal uses a dynamic programing based architecture.

Rather than using PWM or Gibbs sampling for RBS improvement, Prodigal uses a set of bins with predetermined motifs and spacer length. It also assumes that the sequence uses the SD motif. If evidence shows otherwise, Prodigal looks for the alternative motifs described by the earlier mentioned bins.

2.4 Genome and gene sets

2.4.1 Introduction

There also exists the problem of assessing the performance of a gene prediction program. The reason for this is the lack experimentally verified genes. There exist a few gene sets that are verified, but generally it is not possible to assume that there exist verified genes, and thus one have to rely on gene sets like RefSeq (see § 2.4.3).

2.4.2 GenBank

GenBank is a publically available database of annotated sequence data maintained by National Center for Biotechnology Information (NCBI), and '[...] is specifically intended to be an archive of primary sequence data.' (Mizrachi 2002–) This means that all of the sequences in this database is submitted by the person that has performed the sequencing of a given sequence. NCBI does not curate this data, and thus the curation is the

resposibility of the submitter. Any update can only be performed by the author, or a third party with permission from the author. (Mizrachi 2002–)

Also, as the GenBank database is intended as an archival database, there will be duplicates in the database for some loci, since similar experiment will contain similar results.(Mizrachi 2002–)

2.4.3 RefSeq

The Reference Sequence (REFSEQ) database is another database maintained by NCBI. It is a database consisting of '[...] a curated collection of DNA, RNA and protein sequences [...]' (Mizrachi 2002–), and thus differ from the GenBank database by being a non-redundant collection of sequences. Also, the database contains only entries for organisms which has a sufficient amount of data available. (Mizrachi 2002–)

Another difference is that REFSEQ is a curated database built by NCBI, and thus is able to update the annotations without requiring the permission from the submitter. This means that NCBI can update the data when new data appears, and not have to wait until the original submitter updates the entry. (Mizrachi 2002–)

2.4.4 NCBI Genome

NCBI Entrez Genome (`http://www.ncbi.nlm.nih.gov/genome`) is a database containing the whole genomes, with available information, such as its annotations, for prokaryotes, eukaryotes and viruses. It contains genomes from either the primary databases, such as NCBI GenBank, or curated genomes from NCBI REFSEQ. This means that it both contains genomes that are completly sequenced, and genomes that are in-progress of being sequenced. (*New Entrez Genome Released on November 9, 2011* 2011)

Chapter 3

Materials and methods

IN THIS CHAPTER the methods used for storage and analysis are described. The schema for the database used in this thesis is described in § 3.1. In the next section (§ 3.2), the methods used for retrieving the reference data (§ 3.2.1), and how predictions were generated using the gene prediction programs studied in this thesis (§ 3.2.2), is described.

The second half of this chapter is focused on data analysis. The methods for transforming the prediction data into numerical measures are introduced in § 3.3, while the methods used to analyse the transformed data is described in § 3.3.5.

Source code for the custom scripts/programs used in this thesis is available on `http://folk.uio.no/stianeng/code.zip`.

3.1 Storing the data

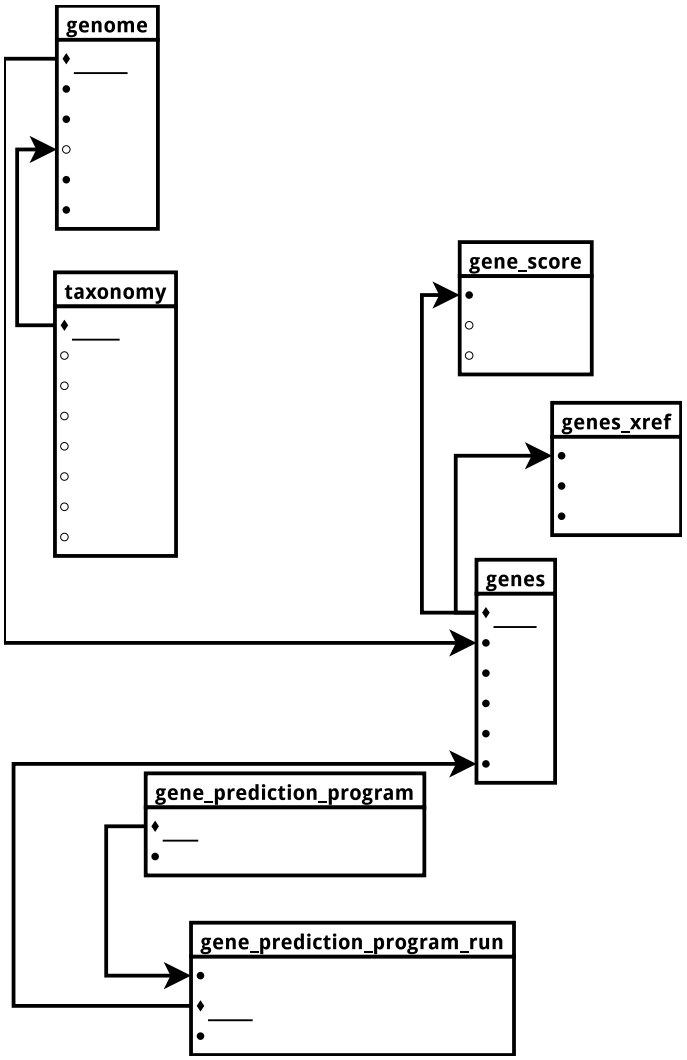
Both the reference data (see § 3.2.1) and the prediction data (see § 3.2.2) were stored in a *relational database.* This allows for easy organization of the data. In addition, a relational database allows for retrieving data using the Structured Query Language (SQL). This makes retrieval of some relevant information without having to write code that processes all the data easy. It also decreses network traffic, by only transferring the desired subset of data. Network traffic is usually is usually slower than local disk access, and thus would drastically reduce the performance.

Due to the availability of a powerful query language, the transformations introduced in § 3.3.3 can be done on the database side. This allows for exploitation of the faster disk access of the database, and thus increases performance.

3.1.1 Schema

The structure of the data needs to be defined using a Data-Definition Language (DDL), since a relational database is used. This structure is called the *schema* of the database, and will be described in the following subsections. See also figure 3.1 for a overview of the database schema and table 3.1 for example data from the database used in this thesis.

Figure 3.1: A diagram showing the database schema used to store the reference data and prediction data used in this thesis.
♦ + underline = Primary key
○ = Can be `NULL`
• = Can not be `NULL`



`genome` The database table `genome` has the role of storing the sequences acquired from NCBI Genome (see section 2.4.4).

The field `gid` has a value that is used as an identifier in this database. Note that this value has no connection to any other external databases.

The field `genbank_accession` holds the identifier that this sequence has in GenBank, and together with the field `genbank_version`, this represents the accession value that a specific sequence has in GenBank.

The field `taxon_id` holds the identifier that this sequnce has in the table `taxonomy`.

The next field, `organism`, hold the organism name for a specific sequence.

The last field `sequence`, is responsible for holding the entire sequence of a given sequence, and will as a result usually be quite large.

`taxonomy` The database table `taxonomy` is responsible for storing the taxonomy data for a given sequence stored in `genome`.

The field `taxon_id` holds the identifier from NCBI Entrez Taxonomy for a given species.

The fields `superkingdom`, `phylum`, `class`, `order_`[1], `family`, `genus` and `species` stores information about the various levels of taxonomy for a given sequence.

`gene_prediction_program` The database table `gene_prediction_program` has the role of storing the identifier, and name of a given gene prediction program.

In this thesis, GenBank will be considered as a 'gene prediction program' in the context of data storage. This lowers the number of tables.

The field `pid` holds the given gene prediction program identifier and is unique to this thesis.

The other field, `name`, holds the name of a given gene prediction program.

`gene_prediction_program_run` The database table `gene_prediction_program_run` has the role of storing information of a run of a specific gene prediction program.

The first field, `pid`, is a reference to the field with the same name in the table `gene_prediction_program` and is used to connect a given gene prediction program run to a given gene prediction program.

The second field, `run_id`, is the identifier for a specific run of a gene prediction program. This is unique to this thesis.

The third field, `parameters`, is the parameters passed to a given gene prediction program for a given run.

[1] `order` is a reserved keyword in SQL

Table 3.1: Example data from the database for the various database tables used in this thesis.

(a) genome

gid	genbank_accession	genbank_version	taxon_id	organism	sequence
506	,NC_000958,	1	243230	,Deinococcus radiodurans R1,	,ATTTTGACC...,(177.5 kb)
507	,NC_000959,	1	243230	,Deinococcus radiodurans R1,	,CCCAGGGCA...,(35.7 kb)
508	,NC_001263,	1	243230	,Deinococcus radiodurans R1,	,TCACGCGAA...,(2648.6 kb)
509	,NC_001264,	1	243230	,Deinococcus radiodurans R1,	,TCTTTGCTC...,(412.3 kb)

(b) taxonomy

taxon_id	superkingdom	phylum	class	order_	family	genus	species
243230	,Bacteria,	,Deinococcus-Thermus,	,Deinococci,	,Deinococcales'	,Deinococcaceae,	,Deinococcus,	,Deinococcus radiodurans,

(c) gene_prediction_program

pid	name
1	,GenBank,
3	,Glimmer,

(d) gene_prediction_program_run

pid	run_id	parameters
1	1	,,
3	3	,,

(e) genes

gene_id	gid	left_end	right_end	strand	run_id
959452	508	1096951	1098030	-1	1
12604945	508	1096951	1098021	-1	3

(f) genes_xref

gene_id	key	value
959452	,GeneID,	,1799966,

(g) gene_score

gene_id	score_type	score
12604945	,RawScore,	16.56

`genes` The role of this table is to store a given gene produced by a given gene prediction program (or gene set).

The first field, `gene_id`, is an identifier for a given gene. This is unique to this thesis.

The next field, `gid`, is a reference to the field with the same name in the table `genome`, and is used to connect a given gene with the sequence that it belongs to.

The two next fields, `left_end` and `right_end`, is responsible to store the left and right coordinates of a given gene in the sequence from the `genome` table.

The next field, `strand`, stores the strand information for a given gene.

The next field, `run_id`, is a reference to the field with the same name in the table `gene_prediction_program_run`, and is used to connect a given gene to the gene prediction program run that it was produced.

`genes_xref` The role of the database table `genes_xref`, is to store any external references that a given gene has, e.g., an entry in EcoGene (Rudd 2000).

The first field, `gene_id`, is a reference to the field with the same name in the table `genes`, and is used to connect a given external gene reference to a given gene.

The second field, `key`, holds the name of the database the external reference is made to.

The third field, `value`, holds the identifier in the external database.

`gene_score` The role of the database table, `gene_score`, is to store scores associated with a given gene, if available.

The first field, `gene_id`, is a reference to the field with the same name in the table `genes`, and is used to connect a given gene with a given score.

The second field `score_type` stores the type of the score, is of type `GENE_SCORE_TYPE`, and contains one of the possible score types found in table 3.2. These values might be used during an analysis of the prediction results. However, these scores were not used during the analysis preformed in this thesis and is included for the sake of completeness.

The last field, `score`, is responsible for storing the value of a given score.

Table 3.2: Possible values for the type GENE_SCORE_TYPE

Score type	Description
,TotalScore,	Prodigal total score
,CodingPotential,	Prodigal coding potential
,StartScore,	Prodigal start score
,RBSScore,	Prodigal RBS score
,UpstreamScore,	Prodigal upstream score
,TypeScore,	Prodigal type score
,AvgProb,	GeneMark average probability
,StartProb,	GeneMark start probability
,RawScore,	Glimmer raw score

3.2 Obtaining and generating the data

3.2.1 Obtaining the reference data

An evaluation of gene prediction programs requires a reference data set to which comparisons can be made. The reference source used in this thesis is NCBI Entrez Genome (see § 2.4.4).

NCBI Entrez Genome contains genomes and their annotations for organisms that have been sequenced. In this thesis, the focus is on the completed prokaryotic sequences (REFSEQ genomes) found at `ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria`. The sequences were retrieved as a gzipped tarball containing the sequences and their annotations in GenBank format on 26 January 2011. It contained 1,262 genomes (2,338 sequences) for 902 different species. These 902 species consisted of 818 bacteria, 84 archaea and 3 bacteriophages.

The NCBI Entrez Taxonomy database was also obtained by downloading it from the ftp page found at `ftp://ftp.ncbi.nih.gov/pub/taxonomy`. This database contains a curated taxonomy of all species registered in NCBI Entrez.

The sequences, with their annotations and their taxonomy were added to the database using a custom script.

3.2.2 Generating prediction data

Gene predictions are performed using a set of custom scripts that first populates a directory structure with fasta files and then uses each prediction program on each fasta file.

The fasta dump process populates a given directory with files on the form `<translation table>/<gid>/<gid>.fa`, where `<translation table>` is the translation table used by the sequence, which in the prokaryotic case is either 11 or 4. `<gid>` is the identifier used by the database for identifying a specific sequence (see § 3.1.1).

After creating a fasta dump, the scripts can be used to run the programs. How the programs were run will be described next. For all programs,

<translation table> and <gid> is assumed to be defined as above.

Glimmer The Glimmer distribution contains several programs, but only four of them are needed for normal usage of Glimmer.

The various programs in the Glimmer distribution were run similarly to how it is run the g3-from-scratch.csh script that comes with the Glimmer distribution. The following commands were run:

- long-orfs -z <translation table> -n -t 1.15 <gid>.fa \
  <gid>.longorfs
- extract -t <gid>.fa <gid>.longorfs > <gid>.train
- build-icm -r <gid>.icm < <gid>.train
- glimmer3 -z <translation table> -o50 -g110 -t30 <gid>.fa \
  <gid>.icm <gid>

The main output of glimmer3 is a file called <gid>.predict, which contains a list of predicted genes where each line specifies a ORF name, gene start, gene end, strand, reading frame and a *raw score*.

glimmer3 also outputs a file name <gid>.detail containing more detailed information. Note that this file was not used in this thesis.

GeneMark.hmm GeneMark.hmm was run by using the script gmhmmp_heuristic.pl from the GeneMark distribution. This script calculates the GC content of a sequence, and picks a suitable heuristic model among the models in the GeneMark distribution. For each sequence, this script was run as follows:

- gmhmmp_heuristic.pl -gcode <translation table> -outfile \
  <gid>.lst -s <gid.fa>

The program outputs a file named <gid>.lst which contains a list of predicted genes where each line contains gene number, strand, left end, right end, gene length, and gene class (?).

GeneMarkS For each sequence, GeneMarkS was run as follows:

- gmsn.pl -gcode <translation table> <gid.fa>

Since this program is based on GeneMark.hmm, the output is on the same form as for GeneMark.hmm.

MED The MED distribution consists of the programs MED2 and TISModel, but only the first program is meant to by run by the user since this programs calls the other program internally. The MED2 program does not take any options, which means the translation table cannot be changed, and thus the translation table 11 was used for all sequences.

For each sequence, MED was run as follows:

- ./MED2 <gid.fa>

The output of MED2 is a file containing a list of predicted genes, where each line consists of left end, right end and strand of a predicted gene.

Prodigal The Prodigal distribution only consists of one program, `prodigal`. This program can either be run in a one-step mode, where the program both trains and predicts on the same sequence, or a two-step mode, where training is done on the given sequences, and prediction is done on a possibly different sequences using the models produced in the training step.

Since Prodigal is unable to self-train on sequences shorter that 20 kb, the metagenomic models were used for these sequences in this thesis. This allows Prodigal to produce some result for these sequences.

Thus, for each sequence shorter than 20 kb, Prodigal was run with the command:

- `prodigal -g <translation table> -i <gid>.fa -o <gid>.gbk \ -s <gid>.start -p meta`

Sequences longer or equal to 20 kb, were run in self-training (one-step) mode, i.e., for each sequence, the following command was run:

- `prodigal -g <translation table> -i <gid>.fa -o <gid>.gbk \ -s <gid>.start`

Prodigal will then output a GenBank file file for each sequence, containing the predicted genes. The `prodigal` program is also asked to produce a *start file*[2], which contains all *potential* genes — also those selected as non-coding — with their associated scores. This file was not used in this thesis.

3.2.3 Adding the predictions to the database

All the predictions produced by the gene prediction programs were added to the database using custom programs. These programs extracts the predicted genes and scores for the predicted genes when available.

3.3 Transforming prediction data using measures

The prediction results needs to be transformed using some measures suitable for visualisation and statistical analysis, in order to analyse the accuracy of the gene prediction programs. Different measures are defined in order to evaluate different aspects of the performance of the gene prediction programs.

When compairing a set of predicted genes to a set of reference genes, a score, or several scores, is assigned. How this is done is introduced in the following sections.

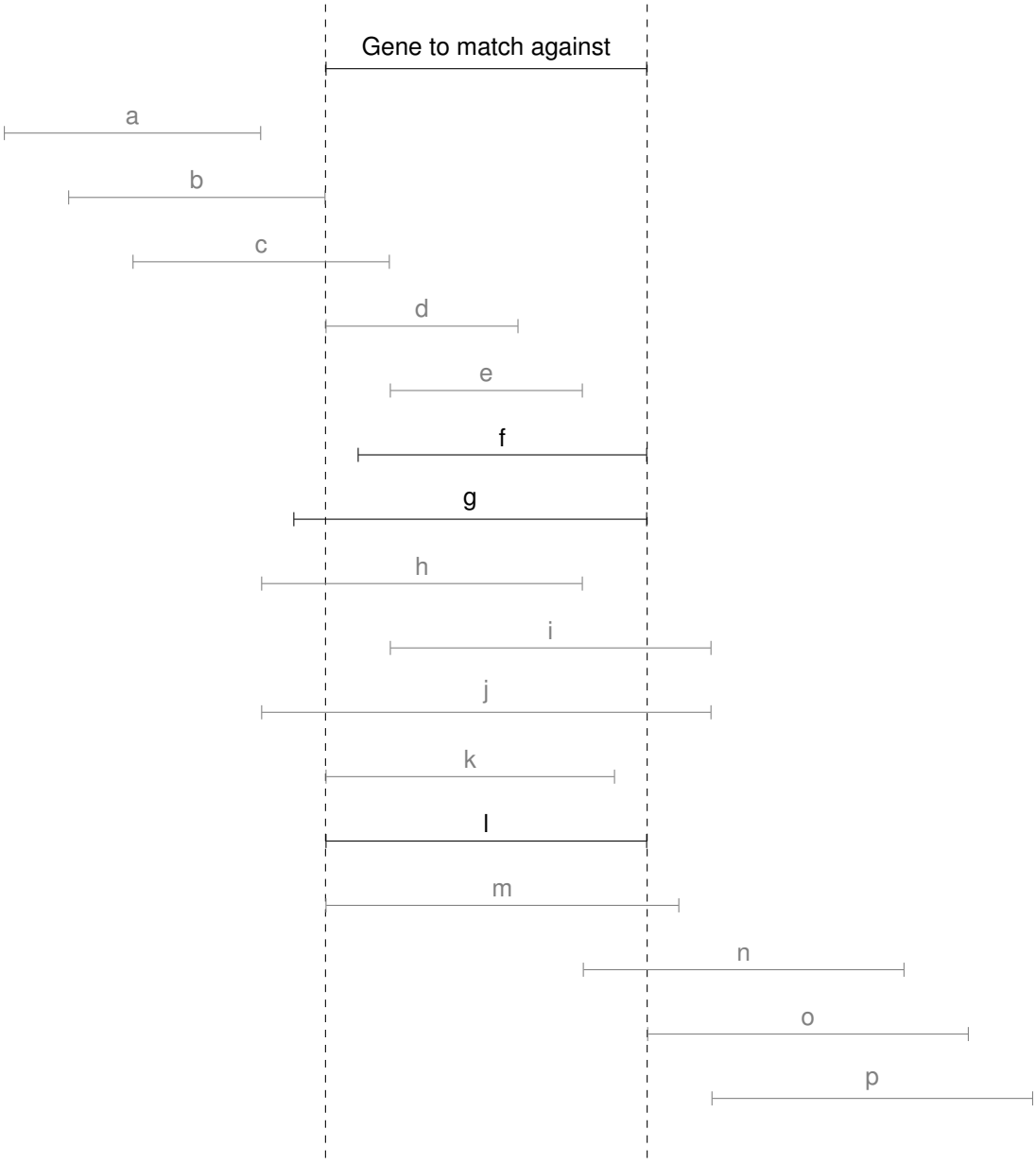
3.3.1 Counting genes

In order to define the measures used for evaluation, it is neccesary to define how to count the genes produced by the gene prediction programs.

[2]Term used by the Prodigal authors

Figure 3.2: Diagram showing different ways to compare predictions.

Genes f, g and l = Exact matching end
Gene l = Exact matching start and end
Genes a–e, h–k and m–p = No exact match



However, there are several ways to count genes. The first way is to count based on matching gene end. This means that, the predicted gene is counted as a match, if the stop coordinates of a given gene matches that of the annotated gene. It is used here since this kind of match were historically widely used for the evaluation gene prediction programs. This kind of match is defined as *gene detection* in this thesis.

The second way is to only count genes based on both matching start and end, i.e., requiring that the start coordinates and stop coordinates matches between a predicted gene and an annotated gene. This kind of match is defined as *gene matching* in this thesis.

The third way is to count genes based on some function, based on the overlap between a predicted gene and an annotated gene. It is possible to say that if some specified function returns a result greater than zero, for a given prediction compared against an annotation, the prediction will be a fuzzy match to the given annotated gene. This was however not used in this thesis.

Various examples of genes that can be counted using one or more of the methods mentioned above are shown in figure 3.2. Predictions f, g and l detect the annotated reference gene. Prediction l correctly matches the annotated reference gene. One or more of predictions c–n are fuzzy matches against the annotated reference gene, depending on how the matching function is defined.

3.3.2 Evaluating predictions

Two ways of counting genes have been defined. Predictions can now be categorized depending on how they compare to a reference gene.

The counts for the various categories is used when scores based on these counts are defined. The categories used for evaluating genes are defined as follows:

Definition 3.1 (True Positive). *True Positive (*TP*) is the number of genes that are equal between a set of predicted genes and a set of annotated genes for the same sequence.* ♣

Definition 3.2 (False Positive). *False Positive (*FP*) is the number of genes that a gene prediction program falsely predicts as correct. This can be expressed as the number of predicted genes minus the number of* TP*s, i.e., if N is the number of predicted genes, then* $\mathrm{FP} = N - \mathrm{TP}$. ♣

Definition 3.3 (False Negative). *False Negative (*FN*) is the number of genes that a gene prediction program falsely discards as incorrect. This can be expressed as the number of correct genes minus the number of* TP*s, i.e., if M is the number of correct genes, then* $\mathrm{FN} = M - \mathrm{TP}$. ♣

However, there are no true numbers for measures such as TP. The numbers have to be estimated by, e.g., using GenBank (see § 2.4.4) as a source for 'correct genes'.

Additionally, in contexts where the categories TP, FP and FN are defined, the category *true negative* is also usually defined. However, since there are no ways to define a true negative gene, true negatives are not defined here.

3.3.3 Evaluating prediction programs

Measures suitable for evaluation of the performance for the various gene prediction programs can now be defined.

Since there are several ways of combining the counts from § 3.3.2, the measures used in this thesis is introduced next.

The measure *recall* is used for the purpose of evaluating the gene prediction programs ability to include true genes.

Definition 3.4 (Recall). *Assume that the measures* TP *and* FN *exists for a given gene prediction program and a given sequence.* Recall *is then defined as*

$$r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3.5}$$

The measure *precision* is used for the purpose of evaluating the gene prediction programs ability to exclude false genes.

Definition 3.6 (Precision). *Assume that the measures* TP *and* FP *exists for a given gene prediction program and a given sequence.* Precision *is then defined as*

$$p = \begin{cases} \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, & \text{if } \mathrm{TP} + \mathrm{FP} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.7}$$

When a program fails to run on a given sequence, TP and FP is zero. As a result of this, precision would be equal $0/0$, and thus undefined. In order to also have results for these sequences, precision is defined — in this thesis — in such a way that a failure to run on a given sequence would produce $p = 0$.

Precision and recall can be combined into a weighted score known as the F_β-score. However, this was not used in this thesis.

3.3.4 Visualisation of the data set

Most of the graphics were created using the R (R Development Core Team 2011) package ggplot2 (Wickham 2009).

Genes per megabase The number of annotated genes were plotted against the sequence length, in order to check if the number of genes are dependent on sequence length. In addition, a regression line for the model #CDS ~ length was plotted.

A similar plot was also, for all of the gene prediction programs, created for the number of predicted genes against the sequence length. The purpose was to check whether the number of predicted genes showed a dependece on the sequence length.

Violin plots *Violin plots* were used in order to visualise precision and recall for all of the gene prediction programs. A violin plot combines a box plot with a density plot (Hintze and Nelson 1998).

The violin plots were produced using the functions `geom_violin` and `geom_boxplot` from the R package ggplot2.

Violin plots were also used to visualise the residuals from the lasso models (described later in § 3.3.5).

Start and stop codon difference Plots showing the difference between annotated and predicted start codon were created to assess the gene prediction programs' ability to accurately match detected genes.

Assume that X is the set of predicted genes, and Y is the annotated genes. The genes in X and Y are on the form

$$z = (l, r, s, g), \tag{3.8}$$

where l is the left end of the gene, r is the right end of the gene, s is the strand and g is the sequence identifier (`gid`).

This can be transformed into a form using gene start and end instead of left end, right end and strand. This can be done by defining a function

$$h(z) = h(l, r, s, g) = \begin{cases} (l, r, g), & s = + \\ (r, l, g), & s = - \end{cases} \tag{3.9}$$

where $s = +$ denotes the direct strand, and $s = -$ denotes the reverse strand.

Using this function, define the transformed genes as

$$\mathsf{X}' = \{h(a) \mid a \in \mathsf{X}\}, \quad \mathsf{Y}' = \{h(b) \mid b \in \mathsf{Y}\}. \tag{3.10}$$

The differences between annotated and predicted start and stop codons are then defined as

$$\Delta = \{(a - x, b - y, g) \mid (x, y, g) \in \mathsf{X}', (a, b, g') \in \mathsf{Y}', g = g'\}. \tag{3.11}$$

For convenience, define $\Delta_{5'}$ to denote the difference between annotated start codon and predicted start codon. Similarly, define $\Delta_{3'}$ to denote the difference between annotated stop codon and predicted stop codon. Using these, it is possible to define the probability for a start codon difference x given stop codon difference y and sequence z as

$$\mathrm{P}(\Delta_{5'} = x \mid \Delta_{3'} = y, g = z) = \frac{\#\{(a, b, c) \mid (a, b, c) \in \Delta, a = x, b = y, c = z\}}{\#\{(a, b, c) \mid (a, b, c) \in \Delta, b = y, c = z\}}. \tag{3.12}$$

Similarly, the probability for a stop codon difference x given start codon difference y and sequence z can be defined as

$$\mathrm{P}(\Delta_{3'} = x \mid \Delta_{5'} = y, g = z) = \frac{\#\{(a, b, c) \mid (a, b, c) \in \Delta, a = y, b = x, c = z\}}{\#\{(a, b, c) \mid (a, b, c) \in \Delta, a = y, c = z\}}. \tag{3.13}$$

With these, define the probability that a detected gene is matched given a sequence z can be defined as

$$\mathrm{P}(\Delta_{5'} = 0 \mid \Delta_{3'} = 0, g = z). \tag{3.14}$$

Similarly, the probability that a gene with correctly predicted start is matched for a given sequence z is defined as

$$\mathrm{P}(\Delta_{3'} = 0 \mid \Delta_{5'} = 0, g = z). \tag{3.15}$$

'Continuous box plots' were constructed for both measures by calculating the percentiles $p_5, p_{25}, p_{50}, p_{75}, p_{95}$ at each position x for the observed distributions of

$$\mathrm{P}(\Delta_{5'} = x \mid \Delta_{3'} = 0) \tag{3.16}$$

and

$$\mathrm{P}(\Delta_{3'} = x \mid \Delta_{5'} = 0). \tag{3.17}$$

These percentiles was plotted as shaded regions with different levels of opacity for the regions between various percentiles. The p_{50} percentile were plotted with 0% opacity, the regions p_{25} to p_{50} and p_{50} to p_{75} were plotted with 20% opacity, and the regions p_5 to p_{25} and p_{75} to p_{95} were plotted with 60% opacity.

3.3.5 Statistical analysis of the data set

Linear regression was used to explain the results produced by the various gene prediction programs. Precision and recall were used as the response variables, while GC content, sequence length, genetic code and taxonomy were used as the explanatory variables. The model analysed was thus

$$\begin{aligned} Y = \beta_0 &+ \beta_1 \cdot \mathsf{GC} + \beta_2 \cdot \mathsf{length} \\ &+ \beta_3 \cdot \mathsf{genetic\ code} + \boldsymbol{\beta} \cdot \mathbf{factor}(\mathsf{taxonomy}) + \varepsilon, \\ &\varepsilon \sim \mathrm{N}(0, \sigma^2), \quad \boldsymbol{\beta} = (\beta_4, \beta_5, \ldots, \beta_{1742}) \in \mathbb{R}^{1738} \end{aligned} \tag{3.18}$$

where Y is either precision or recall, GC is the GC content in percent, length the sequence length in nucleotides, genetic code is the genetic code and σ^2 the variance of the noise. The vector $\mathbf{factor}(\mathsf{taxonomy}) \in \{0,1\}^{1738}$ contains the indicator functions for taxonomy at the levels (i) superkingdom; (ii) phylum; (iii) class; (iv) order; (v) family; (vi) genus; and (vii) species. The factor thus takes the form

$$\begin{aligned} &\mathbf{factor}(\mathsf{taxonomy}) \\ &\quad = (\mathrm{I}(\mathsf{domain} = \mathsf{Bacteria}), \ldots, \mathrm{I}(\mathsf{phylum} = \mathsf{Cyanobacteria}), \ldots). \end{aligned} \tag{3.19}$$

This model is however unsuitable for further analysis since the number of coefficients is too large. Since there are only 2,338 sequence in the reference data set used, there will only be 2,338 data points per gene prediction program.

Ordinary least squares is unsuited in this case, because the large number of coefficients means that any estimates will be based on a low number of data points. If the data is assumed to be evenly spread and non-singular, which is unrealistic in this case, there are on average 1.34 data points for each coefficient in the model. Thus the coefficients may be poorly determined, and as a result exhibit high variance.

This motivates the usage of statistical methods that reduces the number of coefficients.

Reducing the number of coefficients

This section was based on Hastie, Tibshirani and Friedman (2009).

It is possible to end up with a small enough number of coefficients by reducing the large number of explanatory variables present in the full model (3.18).

Using *subset selection* is a simple way of achieving a model with a smaller number of explanatory variables. In subset selection, the complexity of the model is reduced by only keeping a small number of the explanatory variables. After using a suitable strategy for subset selection, ordinary linear regression is performed on the subset of coefficients chosen by the subset selection method.

While performing subset selection will help getting a model with higher interpretability than the full model, it is still a discrete process. This can lead to a high degree of variance, and as a result not be able to reduce the prediction error of the model. By using *shrinkage methods* instead of subset selection, it is possible to reduce this prediction error. This is because the shrinkage methods are continuous, rather than discrete, and will thus suffer less from variability.

The two most commonly used shrinkage methods are *ridge regression* and the *lasso*. In both methods, a restriction of the size of the coefficients in the regular least squares estimates are imposed. The two methods differ by how coefficients are restricted. The ridge regression estimate is defined as

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\big(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2\big), \tag{3.20}$$

while the lasso estimate is defined as

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\big(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\big). \tag{3.21}$$

In both cases, $\boldsymbol{y}$ is the response variables, $\boldsymbol{X}$ the input matrix containing the centred explanatory variables, $\boldsymbol{\beta}$ the vector of the coefficients (without the intercept), $\|\cdot\|_p$ the ℓ_p-norm and λ is the parameter controlling the amount of shrinkage.

While a closed form solution of (3.20) for the ridge regression exists, no closed form solution of (3.21) for the lasso exists. Methods which solves the lasso with the same computational complexity as the ridge regression have however been developed.

The lasso was here chosen for its ability to set coefficients exactly equal to zero. A list of coefficients for closer study can then be made by filtering out the non-zero coefficients.

Performing the lasso

The lasso estimates were found using the R package glmnet (Friedman, Hastie and Tibshirani 2010).

After performing the lasso, the model with the least number of non-zero coefficients within one standard deviation of the model with the lowest cross-validation[3] error, was chosen. This is commonly known as the 'one-standard-error' rule. (Friedman, Hastie and Tibshirani 2010)

[3] 10 fold cross-validation is the default used by glmnet

Chapter 4

Results

In this chapter the results of using the methods introduced in chapter 3 is described. The results are spread over two sections.
The first half of the chapter (§ 4.1) contains various plots of the data visualised using the methods introduced in § 3.3.4. The second part of the chapter (§ 4.2) contains the results of performing the statistical analysis using the methods introduced in § 3.3.5.

4.1 Visualisation of the data

4.1.1 Number of genes v. sequence length

Linear regression was performed on the model $\#\mathsf{CDS} \sim \beta \cdot \mathsf{length\ (Mb)}$, where #CDS denotes the number of coding genes, to check if the number of coding genes dependents on the sequence length. This gave a estimated slope of $\hat{\beta} = 896.46 \approx 896$ with $r^2 = 0.99$. This is represented by a grey line in figure 4.1. It is also represented in figure 4.2a as a blue line. This shows that the number of coding genes appears to be only dependent on the length of the sequence.

A similar regression was also performed for the number of predicted genes for each of the gene prediction program to see if the programs follow the same trend as the annotations. For each program, a linear regression on the model $\mathsf{TP} + \mathsf{FP} \sim \beta \cdot \mathsf{length\ (Mb)}$ was performed. This gave estimated slopes as listed in figure 4.2c and is also shown as lines in figure 4.2a.

It is noticable that the r^2 for all programs are 0.90 or larger and given the performance of these fitted models, for all programs, it seems that the number of predicted coding genes are only dependent on the sequence length.

A 95% confidence interval for the slope β was also constructed for each program. Looking at these confidence intervals in figure 4.2b the programs fall into three groups: (i) GenBank, GeneMarkS, Prodigal;[1] (ii) GeneMark.hmm, Glimmer; and (iii) MED.

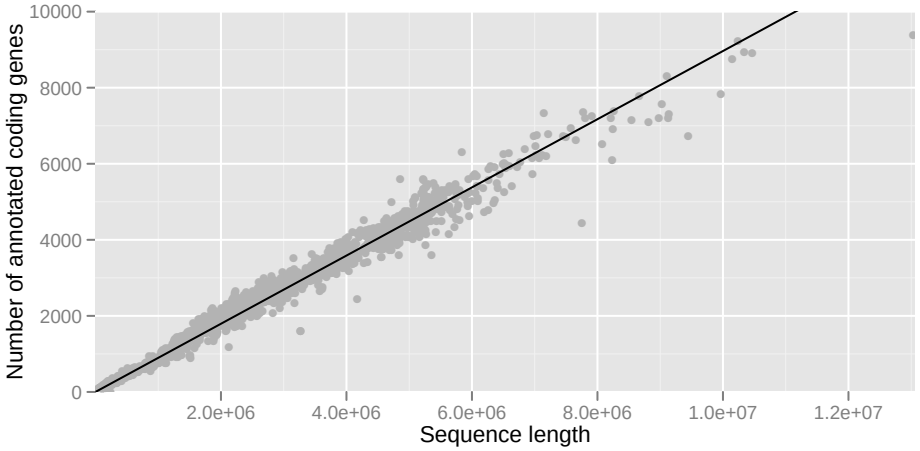
[1] Although the estimated slope for Prodigal is significantly different from GenBank, neither are significantly different from GeneMarkS.

Table 4.1: Table showing the pseudogene rate (pgr), number of annotated coding genes per megabase (cl), the number of predicted coding genes per megabase (pl), recall (r) and precision (p) for selected species.
The numbers are the averages of all sequences for a given species.

					$5' + 3'$		$3'$	
Species	pgr	cl	Program	pl	r	p	r	p
Anabaena azollae	45.7%	546	GeneMark.hmm	1013	66.1%	39.2%	82.1%	49.2%
Phylum: *Cyanobacteria*			GeneMarkS	1043	60.0%	37.3%	79.5%	48.2%
GC: 35.9%			Glimmer	1584	76.3%	29.2%	93.3%	35.4%
			MED	1412	42.8%	46.5%	63.3%	52.5%
			Prodigal	1111	62.4%	33.4%	82.0%	49.3%
Mycobacterium leprae	69.5%	491	GeneMark.hmm	927	60.0%	31.8%	94.6%	50.1%
Phylum: *Actinobacteria*			GeneMarkS	901	61.5%	33.5%	95.3%	51.9%
GC: 57.8%			Glimmer	1871	54.3%	14.3%	95.4%	25.0%
			MED	3070	60.3%	9.6%	95.1%	15.2%
			Prodigal	1221	68.9%	27.7%	97.9%	39.9%
Mycobacterium ulcerans	18.8%	602	GeneMark.hmm	1298	48.2%	28.6%	93.3%	49.8%
Phylum: *Actinobacteria*			GeneMarkS	1293	43.7%	27.0%	93.1%	45.6%
GC: 64.1%			Glimmer	640	32.7%	29.8%	78.8%	81.6%
			MED	1047	27.7%	15.9%	95.5%	54.6%
			Prodigal	727	52.1%	43.1%	91.5%	78.2%
Orientia tsutsugamushi	31.7%	767	GeneMark.hmm	1100	78.3%	55.4%	95.1%	67.0%
Phylum: *Proteobacteria (alpha)*			GeneMarkS	1123	77.7%	53.9%	96.1%	66.1%
GC: 30.5%			Glimmer	1137	73.3%	50.5%	95.5%	65.1%
			MED	1228	59.8%	38.5%	93.9%	59.1%
			Prodigal	1120	82.2%	57.2%	97.0%	67.1%
Rickettsia massiliae	42.5%	748	GeneMark.hmm	1160	53.9%	35.0%	95.0%	61.1%
Phylum: *Proteobacteria (alpha)*			GeneMarkS	1148	51.9%	34.0%	95.3%	62.0%
GC: 32.1%			Glimmer	1126	43.2%	29.5%	82.7%	54.1%
			MED	742	18.7%	9.4%	53.7%	74.8%
			Prodigal	1227	53.2%	32.4%	95.2%	58.2%
Sodalis glossinidius	40%	681	GeneMark.hmm	1452	63.1%	30.0%	81.6%	39.0%
Phylum: *Proteobacteria (gamma)*			GeneMarkS	1368	60.5%	30.4%	77.5%	39.4%
GC: 49.0%			Glimmer	1708	63.8%	25.9%	89.8%	36.4%
			MED [a]	914	33.6%	12.0%	45.9%	16.1%
			Prodigal	1628	60.8%	25.5%	77.7%	32.5%
Trichodesmium erythraeum	14%	574	GeneMark.hmm	735	86.4%	67.5%	96.6%	75.5%
Phylum: *Cyanobacteria*			GeneMarkS	683	84.0%	70.7%	96.2%	81.0%
GC: 34.1%			Glimmer	1201	77.1%	36.9%	93.8%	44.8%
			MED [a]	871	73.6%	48.6%	97.1%	64.0%
			Prodigal	701	90.2%	73.9%	96.9%	79.4%

[a] MED failed to run on 50% of the sequences for this species

Figure 4.1: Graph showing the number of annotated coding genes plotted against the sequence length.



Outliers Figure 4.2a shows some outliers outside the fitted line. A further investigation of both the outliers above the line and the outliers below the line, only revealed a pattern for the outliers below the line. Some of these are listed in table 4.1 along with their performance, pseudogene rate, number of annotated coding genes and the number of predicted coding genes.

The pseudogene rate is the ratio between the number of annotated coding genes and the number of pseudogenes. The number of pseudogenes was extracted from the annoatations in the reference data set for all of the listed species except *Sodalis glossinidius*. Since the pseudogenes were not annotated in the reference for *S. glossinidius* (accession `NC_007712.1`), the number of pseudogenes were extracted from Toh et al. (2006).

It is evident from table 4.1 that the performance for precision is lower than the performance for recall. For these outliers it appears that gene matching is harder than gene detection.

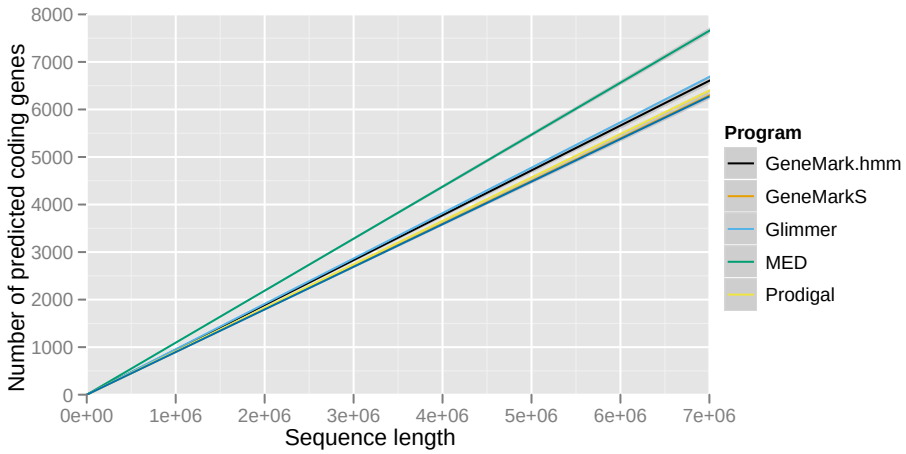
Compairing the various programs, it seems that MED has the lowest performance for these outliers, while GeneMark.hmm and Prodigal seems to have the highest performance.

4.1.2 Violin plots for precision and recall

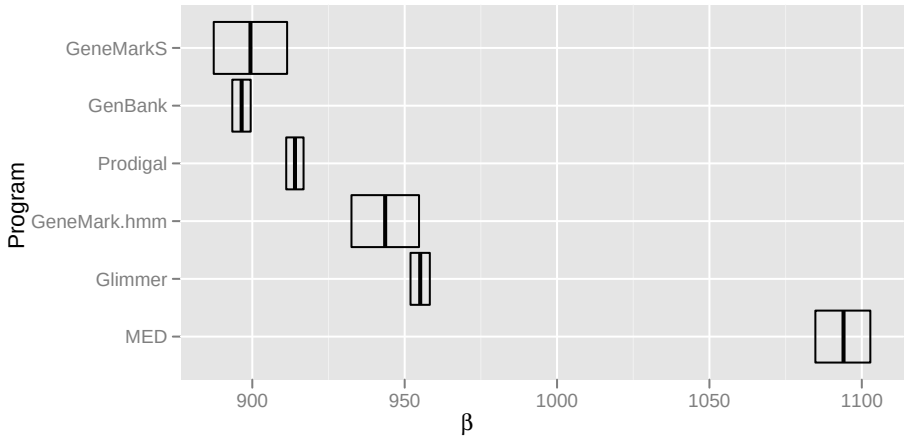
Violin plots were created in order to visualise the performance of the various gene prediction programs for all 2,338 sequences. These violin plots were created for both exact matching gene — to see if the programs are able to find all of the annotated genes — and exact matching gene start and end — to see of they are able to achieve a high degree of accuracy when finding the genes. These plots were created using `geom_violin` and `geom_boxplot` from the R package ggplot2.

Figure 4.2: Graph showing the number of predicted genes plotted against sequence length.

(a) Graph of the number of predicted genes plotted against sequence length.



(b) 95% CI for the slopes in figure 4.2a.



(c) Table showing statistics for figures 4.2a and 4.2b.

Program	Estimated slope	r^2	95% CI
NCBI Genome	896	0.99	893, 899
MED	1094	0.96	1085, 1103
Glimmer	955	0.99	952, 958
GeneMark.hmm	944	0.92	933, 955
GeneMarkS	899	0.90	887, 911
Prodigal	914	0.99	911, 917

Gene detection Looking at the gene detection plot (see figure 4.3) all of the programs are generally able to detect most of the genes. By also looking at the proportion of sequences with recall above 90% (see table 4.2) the programs are generally doing well with regard to gene detection.

However, looking at the precision for the programs, the precision of gene detection is slightly worse. This is especially noticable for MED, which has a proportion of sequences with precision above 90% of 12.9%. Glimmer performs best with regard to recall. However, the performance of Glimmer is second worst with regard to precision.

Lastly, it is also noticable that only MED and GeneMark.hmm achieves the same rank for precision and recall.

Gene matching The gene matching plots (see figure 4.4) reveals that the accurate matching of genes is more difficult than detection, as is expected from the theory (see § 2.1.4). In addition to being harder and thus leading to worse performance, it is also obvious that the performance — for both precision and recall — has a higher degree of spread that the gene detection performance.

When looking at the proportion of sequences with recall above 90% (see table 4.2) the performance is quite a bit lower for gene matching compared with gene detection. However, it is noticable that the rank for recall is mostly the same for both detection and matching for all prediction programs.

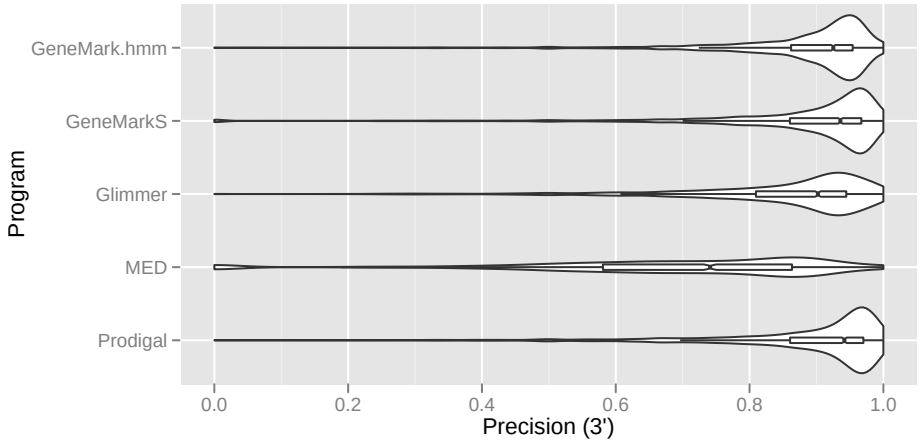
When looking at the precision the programs perform worse for precision when compared to recall. However, Prodigal has the same proportion for both recall and precision. The rank for precision differs for all programs with the exception of GeneMarkS and Prodigal when comparing detection and matching.

Table 4.2: Table showing the proportion of sequences with precision or recall over 90% for both gene detection and gene matching. The rank, based on a given measure, is shown inside parentheses.

	Detection		Matching	
Program	Precision	Recall	Precision	Recall
MED	12.9% (5)	74.6% (5)	1.9% (5)	4.3% (3)
Glimmer	51.6% (4)	88.5% (1)	1.9% (5)	3.1% (5)
GeneMark.hmm	63.1% (3)	81.9% (3)	2.4% (3)	3.8% (4)
GeneMarkS	65.5% (2)	78.3% (4)	5.0% (2)	9.8% (2)
Prodigal	66.4% (1)	82.9% (2)	24.7% (1)	24.7% (1)

Figure 4.3: Violin plots for gene detection.

(a) Precision



(b) Recall

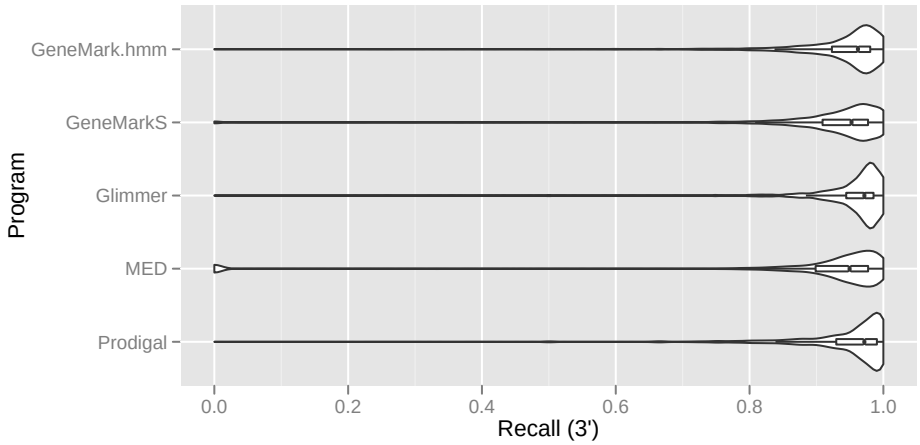
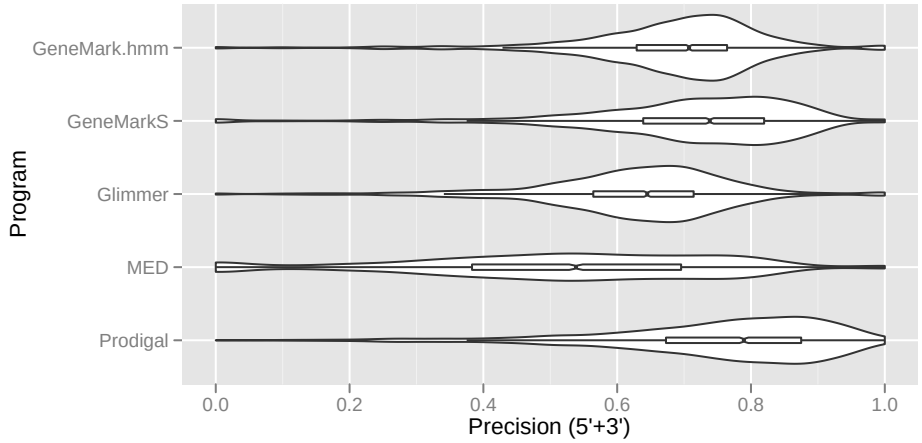
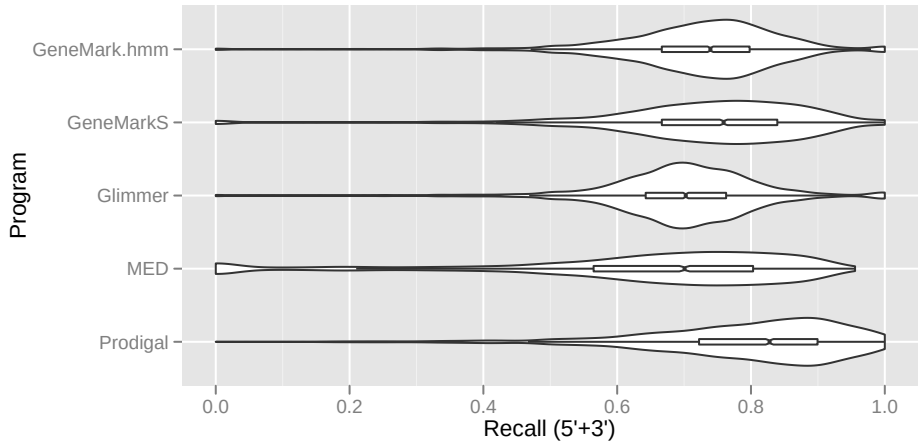


Figure 4.4: Violin plots for gene matching.

(a) Precision



(b) Recall



4.1.3 Accuracy of start and stop codon prediction

A plot showing the difference between annotated and predicted genes (see § 3.3.4) was created to see how well the different gene prediction programs accurately predict the correct start and stop codon for detected genes. Looking at figure 4.5 and table 4.3 the 95-percentile probability that a correct start codon is predicted given a correct stop codon is over 75% for all programs.

It is also clear that most incorrectly predicted start codons have a difference with multiplier 3 ($\Delta = \{0, \pm 3, \pm 6, \ldots\}$), i.e., a difference in whole number of codons. The differences with multiplier 3 also seems to follow a symmetric distribution, and thus the programs are equally likely to overpredict or underpredict the correct start codon.

Few points with multiplier 1 or 2 are found. However, most of these have probability of less than 10^{-4}.

The median probability for correctly predicting the correct start differs by about 0.2 between the program with lowest probability and highest probability. See table 4.3 for more details.

A similar plot for the difference between predicted and annotated stop codon for predicted genes with correct start, were also produced. From figure 4.6 it is evident that the probability for correctly predicting the stop codon given a correctly predicted start codon, is ~ 100% for all gene prediction programs, even for the 5-percentiles. Most of the other points have a probability of less than 10^{-5}.

Lastly, some 'conserved' peaks are present in both figures 4.5 and 4.6.

Figure 4.5: Graphs showing the differences between predicted and annotated start codons for predicted genes with correct stop.
For each difference modulo 3, represented with different colours, the p_{50} percentile are plotted with 0% opacity, the regions p_{25} to p_{50} and p_{50} to p_{75} are plotted with 20% opacity, and the regions p_5 to p_{25} and p_{75} to p_{95} are plotted with 60% opacity.

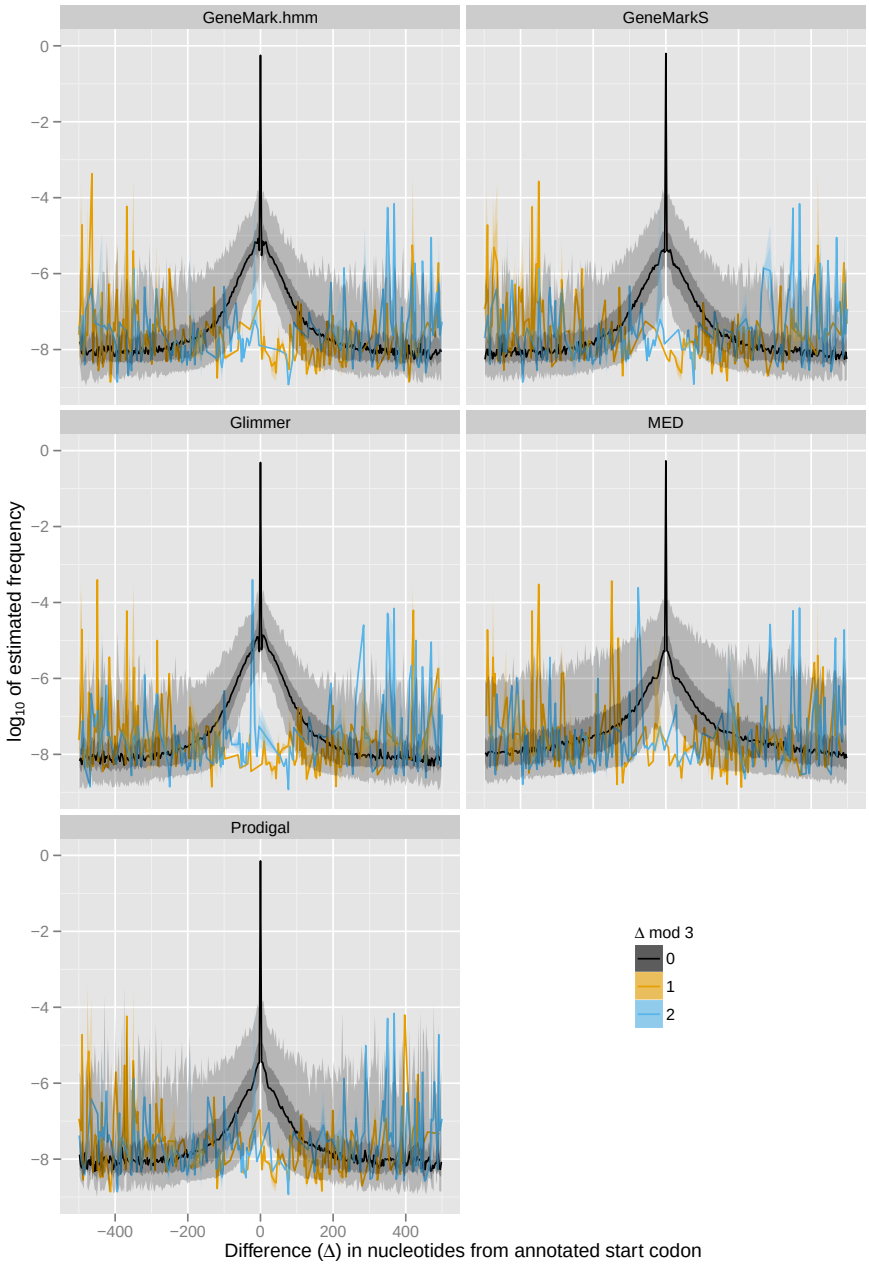


Figure 4.6: Graphs showing the differences between predicted and annotated stop codons for predicted genes with correct start.
For each difference modulo 3, represented with different colours, the p_{50} percentile are plotted with 0% opacity, the regions p_{25} to p_{50} and p_{50} to p_{75} are plotted with 20% opacity, and the regions p_5 to p_{25} and p_{75} to p_{95} are plotted with 60% opacity.

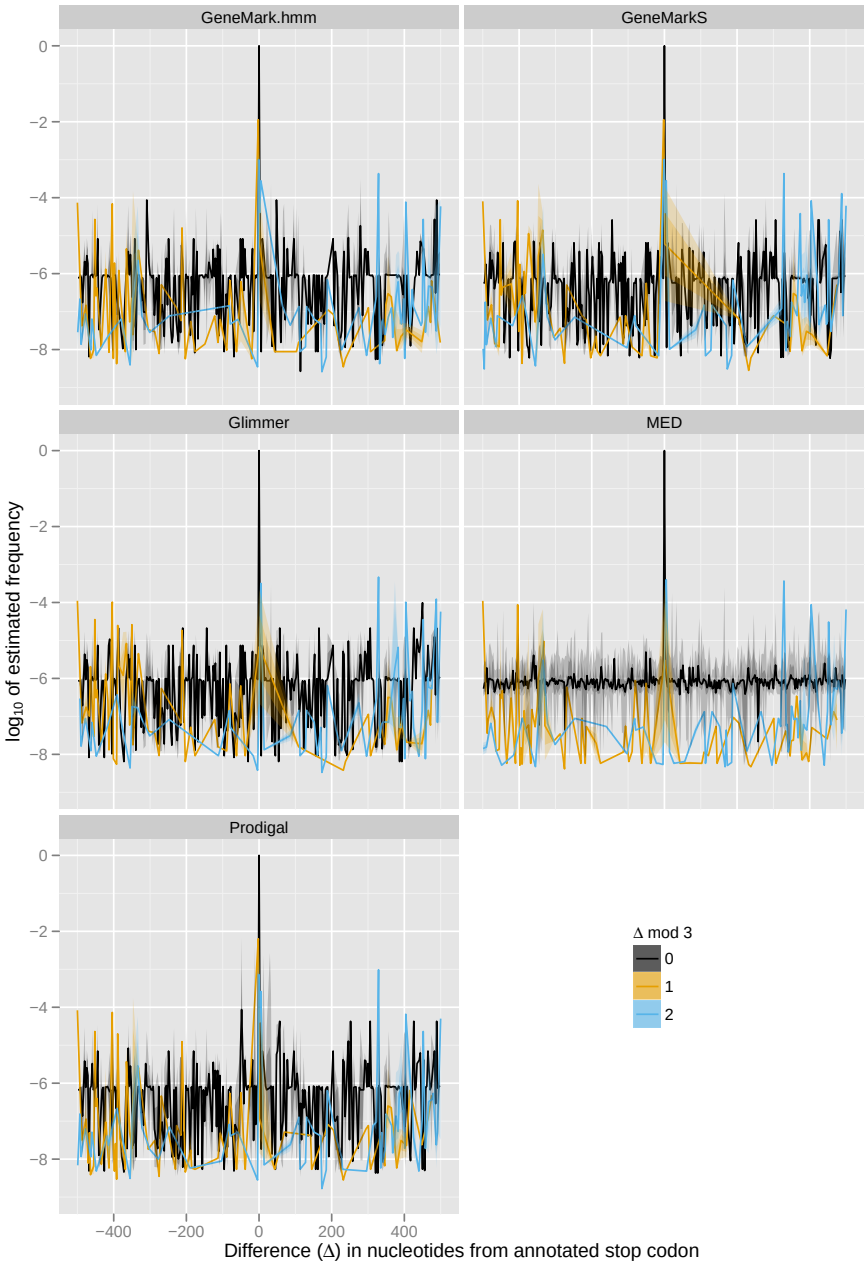


Table 4.3: Table showing some statistics from figures 4.5 and 4.6.
$\Delta_{5'}$ = difference between annotated start codon and predicted start codon.
$\Delta_{3'}$ = difference between annotated stop codon and predicted stop codon.

(a) $P(\Delta_{5'} = 0 \mid \Delta_{3'} = 0)$

	Percentile				
Program	5%	25%	50%	75%	95%
GeneMark.hmm	0.33	0.48	0.56	0.65	0.83
GeneMarkS	0.32	0.50	0.63	0.74	0.89
Glimmer	0.29	0.41	0.49	0.58	0.76
MED	0.15	0.40	0.54	0.69	0.82
Prodigal	0.35	0.57	0.71	0.83	0.96

(b) $P(\Delta_{3'} = 0 \mid \Delta_{5'} = 0)$

	Percentile				
Program	5%	25%	50%	75%	95%
GeneMark.hmm	0.99	1.00	1.00	1.00	1.00
GeneMarkS	0.99	1.00	1.00	1.00	1.00
Glimmer	1.00	1.00	1.00	1.00	1.00
MED	0.99	1.00	1.00	1.00	1.00
Prodigal	0.99	1.00	1.00	1.00	1.00

4.2 Statistical analysis

Since the plots in the previous section reveals that the detection of genes was quite good for all programs, in addition to having a low degree of spread, the gene detection performance was not further analysed.

The gene matching performance, on the other hand, was worse and had higher degree of spread and thus the lasso was performed on the full model, to try to find the cause behind these results.

4.2.1 Full model

The lasso was performed using the R package glmnet (see § 3.3.5). Figure 4.7 shows the CV-plots for the precison models, while figure 4.8 shows the CV-plots for the recall models. The coefficients were chosen using the 'one-standard-error' rule and put into tables. See table 4.3 for precision model coefficients and table 4.4 for the recall model coefficients. The explanatory variables related to taxonomy were put into a taxonomical tree to ease the interpretation of the tables.

CV-plots Looking at the CV-plots and the tables a high degree of shrinkage was applied to the models, for both precision and recall. It seems possible to explain more of the variance for the precision model than the recall model.

In addition, the models for Prodigal explains more than the models for the other programs. This leads to a higher r^2 and lower degree of shrinkage for the models for Prodigal. This means that these models has more coefficients, when compared to the other programs.

Looking at the tables for precision and recall, it is noticable that most of the coefficients are small, but some of the coefficients are larger.

Sequence length The coefficients for sequence length are quite small for both precision and recall. This might indicate that sequence length might only be of importance for separating short sequences that are few kilobases long (plasmids) from longer sequences that are several hundered or thousand kilobases long (megaplasmids, chromosomes).

Translation table The difference in performance between sequences using translation table 11 and sequences using translation table 4 is minuscule.

GC content The coefficients for GC-content are large for all programs. The GC-content coefficient for MED in the precision table is twice as large as the coefficient for the program with the second largest coefficient, Glimmer. However, the coefficient for MED in the recall table is much smaller.

Taxonomy Some large coefficients are scattered around in the taxonomic part of the table. Some examples include the species *Mycobacterium leprae*; *Anabaena azollae*; *Rickettsia massiliae*, the genera *Candidatus Hodgkinia*; *Candidatus Zinderia*; *Sodalis*; *Ureaplasma*, and the orders *Entomoplasmatales*; *Rhizobiales*. For these coefficients the coefficients for MED are zero, or much smaller than the coefficients for the other programs. Some of these examples are also found as outliers in figure 4.1 and also listed in table 4.1.

Intercept Based on the coefficient for the intercept in both tables, it seems that GeneMarkS performs better than Prodigal, followed by GeneMark.hmm, Glimmer and MED. However, since Prodigal has a large number of coefficients, this might not be a clear indication of better performance for GeneMarkS when compared to Prodigal.

Precision v. recall For MED, Glimmer and GeneMark.hmm the coefficients for precision are (slightly) smaller than for recall, while for GeneMarkS and Prodigal the opposite is the case.

Lastly, the recall table has fewer coefficients than the precision table. The coefficients that are common between the precision and recall tables are smaller in the recall table than the precision table.

Residuals From the residuals from the 'one-standard-error' full model — see figure 4.9a for precision and figure 4.12b for recall — it is evident that none of the fitted models have residuals approximately normal distributed. The residuals are also negatively skewed and have a median value significantly larger than zero.

4.2.2 Order model

The lasso were also performed on the order model which is the full model with taxonomy only down to the order level. This was to check if a smaller model that the full model allows for an simpler interpretation of the results, at the expence of the smaller details.

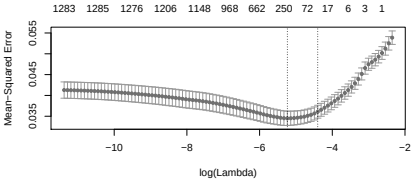
CV-plots Looking at the CV-plots for the order model for precision (see figure 4.10) and recall (see figure 4.11) the order models have a slightly smaller degree of shrinkage than the full models. The number of coefficients and the values of r^2 are also lower for the order models when compared to the full model.

For both order models for GeneMarkS, the 'one-standard-error' model gives only *three* non-zero coefficients. Of these non-zero coefficients, none are part of model that describes the taxonomy.

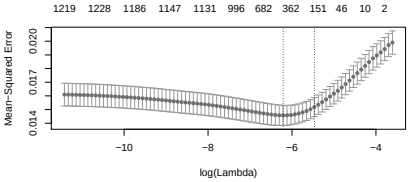
Sequence length Similarly to the full models, the coefficients for sequence length are quite small in the order model for both precision and recall.

Figure 4.7: The CV-plots produced when performing the lasso method on the precision models for the various programs with exact matching gene start and end. In each plot, the left horizontal line represents the model that has the lowest CV-error, while the right line represents the least complex model within one standard deviation of the minimum CV-error. Note that the scales are not equal between the different programs.

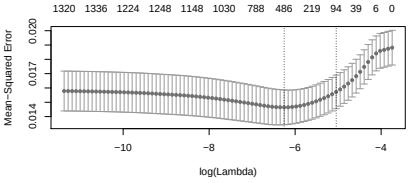
(a) MED



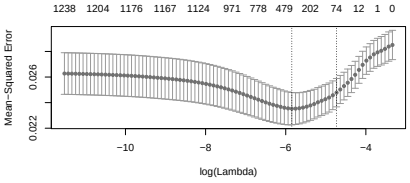
(b) Glimmer



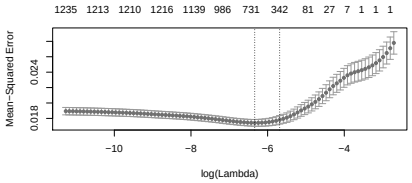
(c) GeneMark.hmm



(d) GeneMarkS



(e) Prodigal

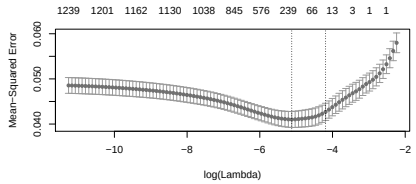


(f) Various measures for the graphs above

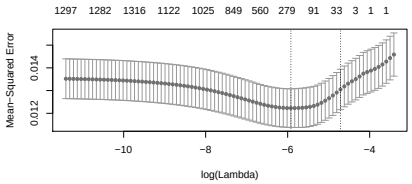
	Lowest CV-error			'One standard error'		
Program	df	r^2	log λ	df	r^2	log λ
MED	219	0.46	-5.24	35	0.35	-4.40
Glimmer	459	0.47	-6.21	175	0.35	-5.47
GeneMark.hmm	486	0.51	-6.26	94	0.28	-5.05
GeneMarkS	382	0.48	-5.85	74	0.26	-4.73
Prodigal	700	0.59	-6.34	342	0.50	-5.68

Figure 4.8: The CV-plots produced when performing the lasso method on the recall models for the various programs with exact matching gene start and end. In each plot, the left horizontal line represents the model that has the lowest CV-error, while the right line represents the least complex model within one standard deviation of the minimum CV-error. Note that the scales are not equal between the different programs.

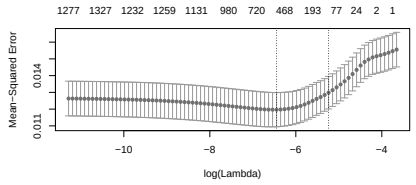
(a) MED



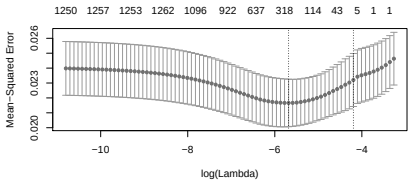
(b) Glimmer



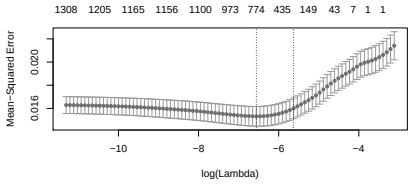
(c) GeneMark.hmm



(d) GeneMarkS



(e) Prodigal

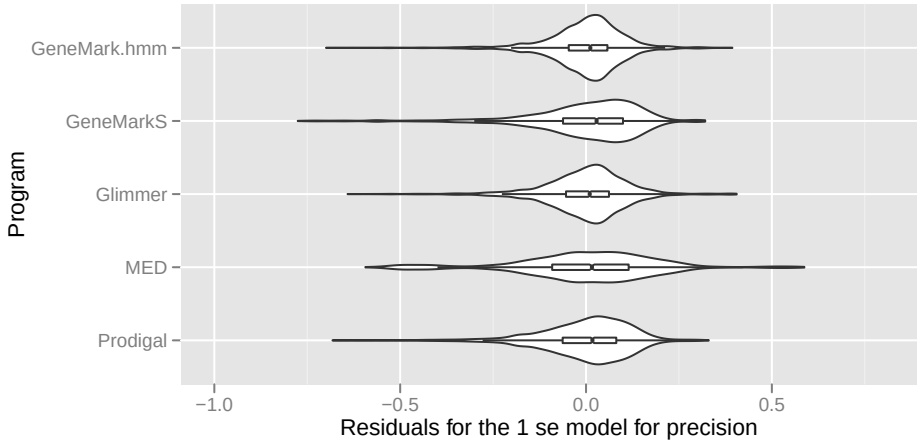


(f) Various measures for the graphs above

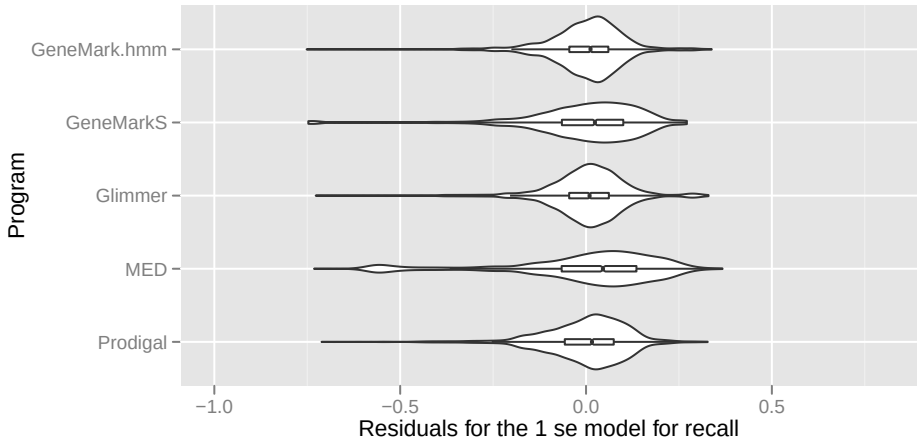
	Lowest CV-error			'One standard error'		
Program	df	r^2	log λ	df	r^2	log λ
MED	203	0.42	-5.12	24	0.29	-4.19
Glimmer	243	0.33	-5.92	22	0.12	-4.71
GeneMark.hmm	552	0.50	-6.45	105	0.28	-5.24
GeneMarkS	287	0.44	-5.68	17	0.09	-4.19
Prodigal	774	0.55	-6.56	299	0.41	-5.63

Figure 4.9: Residuals for the ‘one-standard-error’ full models for precision and recall

(a) Precision



(b) Recall



Translation table The coefficients for the translation table are slightly higher in the order model than in the full model, for those programs which have a non-zero coefficients in both models. It is also clear that the coefficients for the precision order model for Prodigal is quite large (−0.25), when compared to the full model, which has a coefficient of −0.0001, which is $\sim 2,500$[2] times smaller.

GC content Similarly to the full model, the coefficients for GC-content are large for all programs.

Taxonomy Many of the large coefficients that are scattered around in the full models are not found in the order models. However, the order *Entomoplasmatales* — which has large negative coefficients in the full model — are present in the order model with larger coefficients than what is observed in the full model, with the exception of GeneMarkS.

It is also noticeable that MED has large coefficients for this order in the order models, while having these coefficients equal zero in the full model.

Another thing that is noticeable is that 'neighbour' order *Mycoplasmatales* has quite large negative coefficients for MED, while having a small or zero coefficient for the other gene prediction programs.

For the order *Chloroflexalses*, MED has large negative coefficient in both the full models and the order models.

Intercept The coefficients for the intercept in the order models for both precision and recall are close to the estimates for the intercept in the full model. In some cases they are slightly higher and in other cases they are slightly lower.

For the precision order model for Prodigal, the coefficient is almost 0.96. However, Prodigal also has a large negative coefficient (−0.25) for sequences using translation table 11, which is the most common translation table. Thus, the 'base' performance for Prodigal becomes 0.71, which is slightly lower than for the full model.

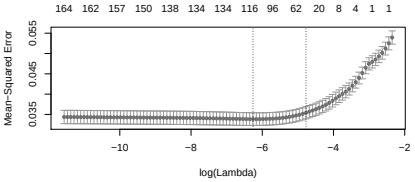
Precision v. recall Similarly to the full model, the recall table has fewer coefficients than the precision table. The coefficients that are common between the precision and recall tables are smaller in the recall table than the precision table.

Residuals From the residuals from the 'one-standard-error' order model — see figure 4.12a for precision and figure 4.12b for recall — it is evident that none of the fitted models have residuals approximately normal distributed. The residuals are also negatively skewed and have a median value significantly larger than zero. This is similar to what is observed for the full model. However, Glimmer has residual distributions that are closer to symmetry than the other programs, when excluding the tails.

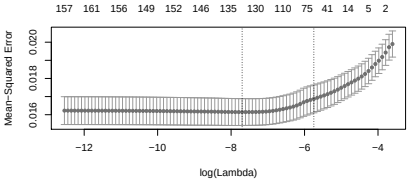
[2] 2,357 when using the coefficients in the tables

Figure 4.10: The CV-plots produced when performing the lasso method on the precision (order) models for the various programs with exact matching gene start and end. In each plot, the left horizontal line represents the model that has the lowest CV-error, while the right line represents the least complex model within one standard deviation of the minimum CV-error. Note that the scales are not equal between the different programs.

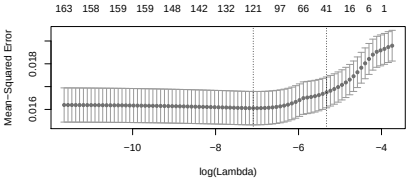
(a) MED



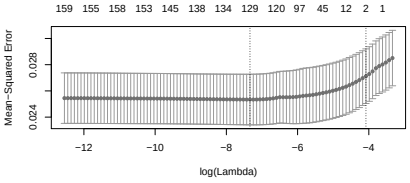
(b) Glimmer



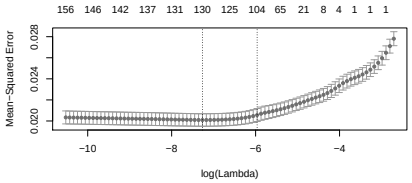
(c) GeneMark.hmm



(d) GeneMarkS



(e) Prodigal

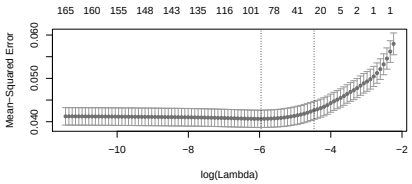


(f) Various measures for the graphs above

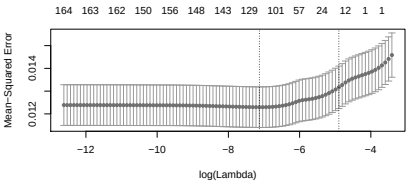
	Lowest CV-error			'One standard error'		
Program	df	r^2	log λ	df	r^2	log λ
MED	111	0.42	-6.26	40	0.36	-4.77
Glimmer	140	0.26	-7.70	70	0.20	-5.75
GeneMark.hmm	121	0.22	-7.09	41	0.14	-5.32
GeneMarkS	129	0.19	-7.34	2	0.05	-4.08
Prodigal	130	0.33	-7.27	104	0.31	-5.96

Figure 4.11: The CV-plots produced when performing the lasso method on the recall (order) models for the various programs with exact matching gene start and end. In each plot, the left horizontal line represents the model that has the lowest CV-error, while the right line represents the least complex model within one standard deviation of the minimum CV-error. Note that the scales are not equal between the different programs.

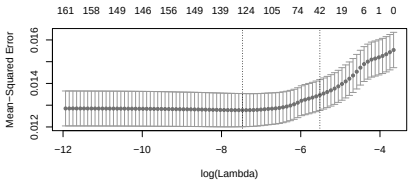
(a) MED



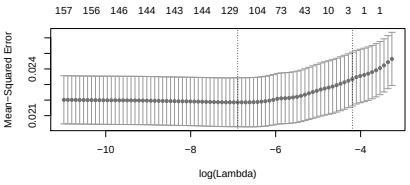
(b) Glimmer



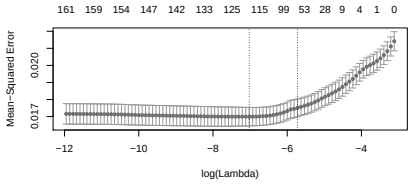
(c) GeneMark.hmm



(d) GeneMarkS



(e) Prodigal

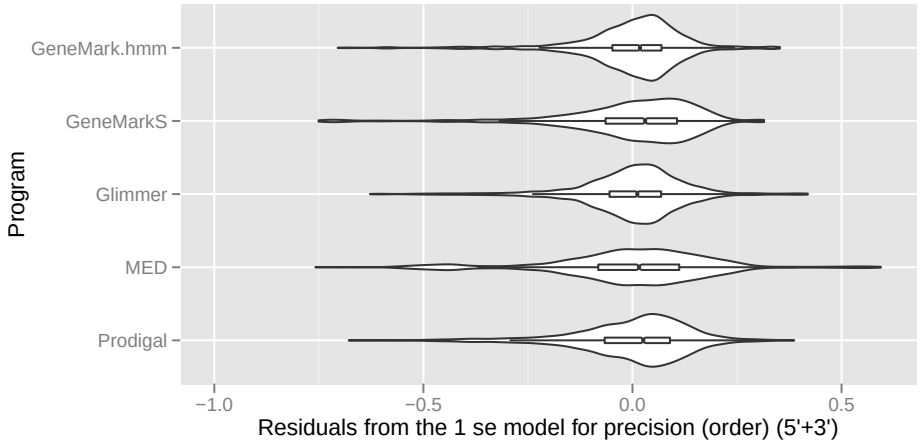


(f) Various measures for the graphs above

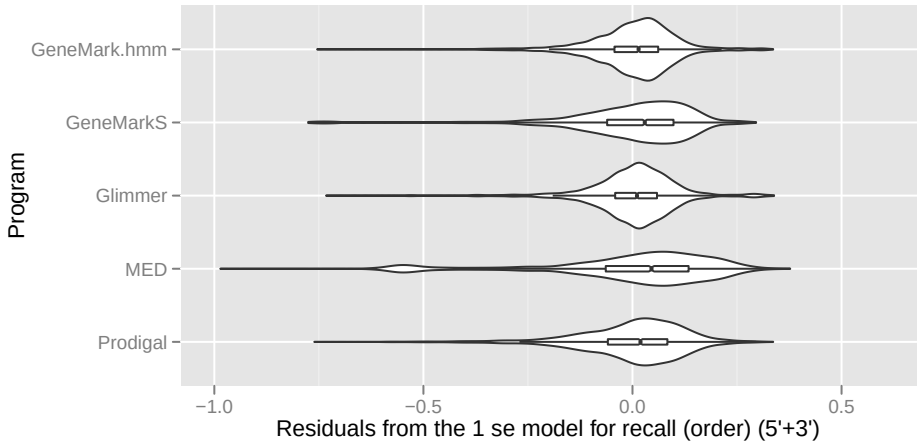
	Lowest CV-error			'One standard error'		
Program	df	r^2	log λ	df	r^2	log λ
MED	93	0.35	-5.96	24	0.29	-4.47
Glimmer	119	0.22	-7.13	13	0.11	-4.90
GeneMark.hmm	134	0.25	-7.47	42	0.17	-5.52
GeneMarkS	119	0.19	-6.82	2	0.05	-4.19
Prodigal	124	0.27	-7.02	68	0.22	-5.72

Figure 4.12: Residuals for the 'one-standard-error' order models for precision and recall. Note that the scales are the same as in figure 4.9.

(a) Precision



(b) Recall



4.2.3 Coefficients for precision and recall

The following pages consist of the tables containing the coefficients obtained by performing the lasso for the precision model (table 4.3) and recall model (table 4.4). Due to their size, the pages containing these tables will be unnumbered and typeset using a small fixed-width font. Table 4.3 is seven (7) pages long and table 4.4 is five (5) pages long.

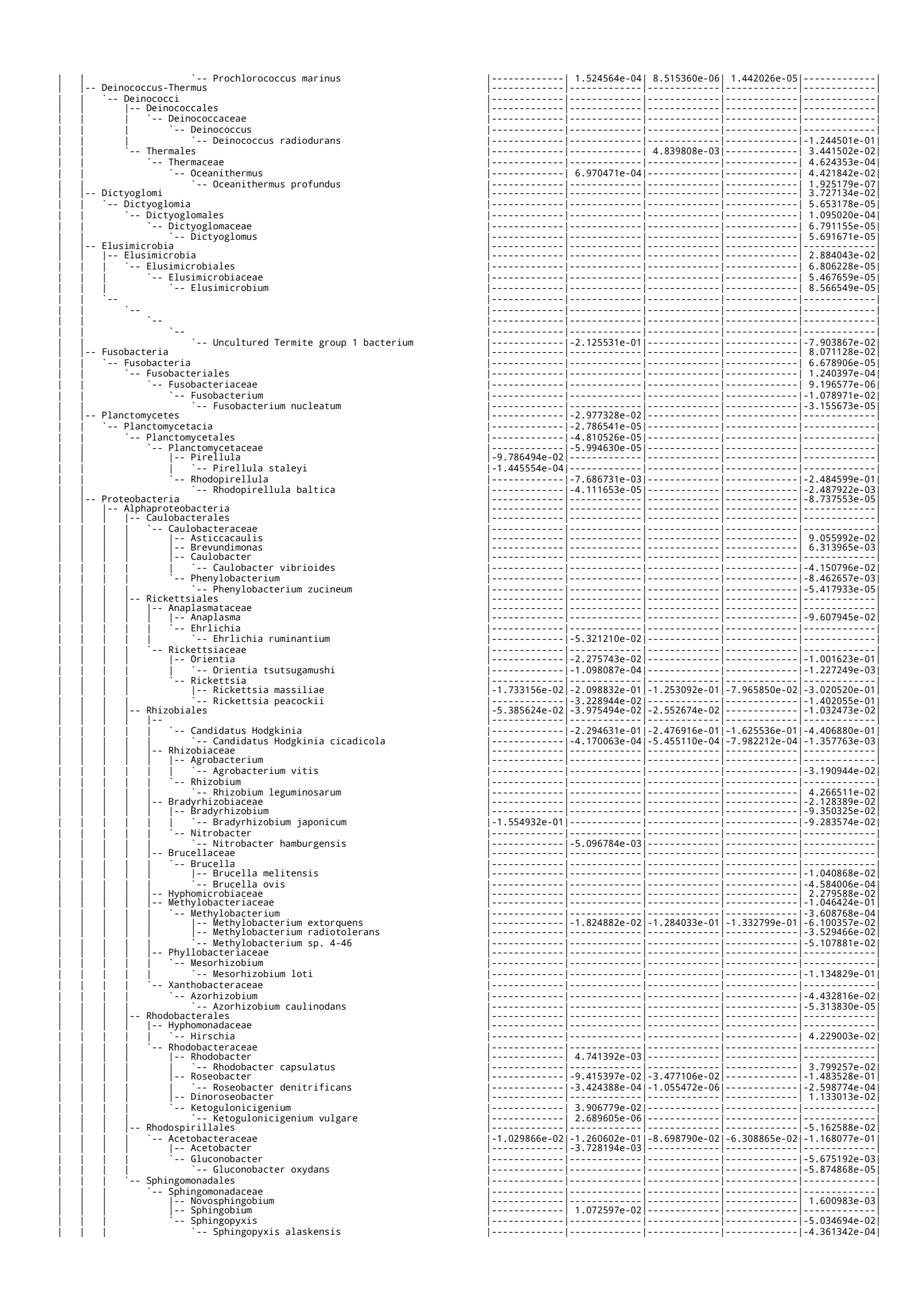
These tables are followed by the tables containing the coefficients obtained by performing the lasso on the order models. See table 4.5 for the precision (order) model and table 4.6 for the recall (order) model. Both tables are two (2) pages long each.

Table 4.3: Lasso coefficients for precision (5'+3')

The table shows the coefficients produced by performing the lasso on the precision model for exact matching gene start and stop. The coefficients for each program are for the model chosen by the 'one-standard-error' rule (see section 3.4.2).
See also figure 4.7 for more information about values of the shrinkage factors.

Coefficients	MED	Glimmmer	GeneMark.hmm	GeneMarkS	Prodigal
Intercept	6.227080e-01	7.199878e-01	7.218945e-01	7.611650e-01	7.389259e-01
GC	-3.872858e-01	-1.886042e-01	-8.628586e-02	-1.604804e-01	-6.274441e-02
Length (bp)	4.703656e-08	1.470067e-08	1.050808e-08	1.641786e-08	3.364753e-08
I(translation table = 11)		-8.301312e-03			-1.045915e-04
Taxonomy					
-- Archaea					
|-- Crenarchaeota					
\`-- Thermoprotei					
|-- Acidilobales		8.075689e-02			
\`-- Acidilobaceae		5.547313e-06			
\`-- Acidilobus					
\`-- Acidilobus saccharovorans		8.838124e-06			
|-- Desulfurococcales					
\`-- Desulfurococcaceae					
|-- Desulfurococcus		1.281037e-01			
\`-- Desulfurococcus kamchatkensis		8.155646e-06			
\`-- Staphylothermus					
\`-- Staphylothermus hellenicus				-3.021010e-01	
\`-- Thermoproteales		3.568368e-04			
\`-- Thermoproteaceae					
\`-- Vulcanisaeta					1.569314e-03
-- Euryarchaeota					
|-- Halobacteria			3.725523e-02		
\`-- Halobacteriales			1.351317e-08		
\`-- Halobacteriaceae					
|-- Halalkalicoccus		-5.660245e-02			-1.577963e-01
\`-- Halalkalicoccus jeotgali		-4.329268e-04			-1.686736e-03
|-- Haloarcula					-7.015630e-02
\`-- Haloarcula marismortui					-2.204489e-03
|-- Halogeometricum		4.315878e-02		1.379046e-02	
\`-- Halogeometricum borinquense				5.705779e-07	
|-- Halomicrobium					7.296101e-03
|-- Haloquadratum					-8.149794e-02
\`-- Haloquadratum walsbyi					-9.377618e-04
|-- Haloterrigena					3.287737e-03
\`-- Natrialba					4.910601e-02
\`-- Natrialba magadii					5.614636e-08
|-- Methanobacteria					
\`-- Methanobacteriales					
\`-- Methanothermaceae					2.932409e-02
|-- Methanothermobacter					-4.375026e-02
\`-- Methanothermus					8.439216e-05
\`-- Methanothermus fervidus					3.324389e-08
|-- Methanococci					1.800808e-02
\`-- Methanococcales					
|-- Methanocaldococcaceae		-6.653369e-03			
\`-- Methanocaldococcus		-3.139116e-05			
|-- Methanocaldococcus infernus					3.073898e-03
|-- Methanocaldococcus jannaschii				-3.420138e-03	-9.234605e-02
\`-- Methanocaldococcus vulcanius		-6.056256e-02			
\`-- Methanococcaceae					1.639032e-02
\`-- Methanococcus					4.335757e-05
|-- Methanomicrobia					
|-- Methanomicrobiales					
\`-- Methanomicrobiaceae					
\`-- Methanoplanus					4.757441e-03
\`-- Methanoplanus petrolearius					1.012855e-07
\`-- Methanosarcinales					
\`-- Methanosarcinaceae					
\`-- Methanosarcina		-1.437037e-01			-1.781755e-01
\`-- Methanosarcina barkeri		-1.335382e-02			-3.548463e-02
|-- Thermococci					
\`-- Thermococcales					
\`-- Thermococcaceae					
\`-- Thermococcus					
\`-- Thermococcus sibiricus		5.069163e-02			
\`--					
\`--					
\`--					
\`-- Aciduliprofundum					3.551480e-02
\`-- Aciduliprofundum boonei					2.621507e-06
\`-- Thaumarchaeota					
\`--					
\`-- Cenarchaeales					-1.505315e-02
\`-- Cenarchaeaceae					-2.843420e-06
\`-- Cenarchaeum					
\`-- Cenarchaeum symbiosum					-1.206817e-06
-- Bacteria		-1.223011e-02	-6.005653e-03		
|-- Acidobacteria					-5.748340e-02
\`-- Solibacteres					-4.154966e-02
\`-- Solibacterales					-5.195737e-06
\`-- Solibacteraceae					-3.011080e-06
\`-- Candidatus Solibacter					-2.063826e-06
\`-- Candidatus Solibacter usitatus					-1.392334e-04
|-- Aquificae		7.867017e-02	1.533554e-02		8.811561e-02
\`-- Aquificae		4.195681e-08	3.387797e-07		5.625676e-05
\`-- Aquificales					2.643855e-04
|-- Aquificaceae					
\`-- Aquifex		-5.214328e-02			-1.740509e-01
\`-- Aquifex aeolicus		-1.844817e-05			-3.873174e-04
\`-- Hydrogenothermaceae		1.945961e-02			
|-- Actinobacteria	-1.764873e-02		-5.384334e-03		
\`-- Actinobacteria	-2.161031e-05		-3.950083e-08		
|-- Actinomycetales		-1.156711e-02	-2.572010e-02	-1.374267e-02	-4.039251e-02
|-- Actinomycetaceae					
|-- Arcanobacterium					1.311364e-02
\`-- Mobiluncus			4.544645e-03		
\`-- Mobiluncus curtisii			3.020490e-05		
|-- Cellulomonadaceae					9.721982e-03
\`-- Cellulomonas					3.791169e-06
|-- Corynebacteriaceae			-5.237088e-03		

				`-- Corynebacterium		-1.726730e-06		
				|-- Corynebacterium aurimucosum				-7.003587e-02
				|-- Corynebacterium efficiens	-2.645284e-03			-2.189388e-01
				`-- Corynebacterium glutamicum				-3.890242e-02
				-- Frankiaceae	-2.869817e-02			-7.446013e-02
				`-- Frankia	-4.007992e-05			-6.540617e-05
				|-- Frankia alni				-1.267931e-01
				`-- Frankia sp. EAN1pec				-1.086137e-02
				-- Microbacteriaceae	-7.841441e-03			-3.052037e-02
				`-- Leifsonia				-7.435169e-02
				`-- Leifsonia xyli				-5.808302e-07
				-- Micrococcaceae				
				`-- Arthrobacter				
				|-- Arthrobacter arilaitensis				-1.754487e-02
				|-- Arthrobacter aurescens				-1.342986e-03
				`-- Arthrobacter chlorophenolicus				8.265186e-02
				-- Mycobacteriaceae		-7.230840e-03	-2.736340e-03	-7.467530e-02
				`-- Mycobacterium		-1.441407e-04	-7.388183e-06	-3.213874e-04
				|-- Mycobacterium abscessus				-1.046513e-02
					-- Mycobacterium leprae	-1.890062e-02	-3.399509e-01	-1.237288e-01	-6.913054e-02	-3.040102e-01
				`-- Mycobacterium ulcerans		-1.677075e-01	-1.465468e-01	-1.188318e-01	-1.336110e-01
			|-- Nocardiaceae					-3.516596e-02
				`-- Rhodococcus		-3.765206e-02	-1.387715e-02		-2.698386e-03
				`-- Rhodococcus opacus		-3.981694e-02			
			|-- Nocardiopsaceae					
				`-- Nocardiopsis					7.115512e-02
				`-- Nocardiopsis dassonvillei					1.305518e-07
			|-- Promicromonosporaceae					7.586390e-02
				`-- Xylanimonas					7.908649e-05
				`-- Xylanimonas cellulosilytica					8.851376e-07
			|-- Pseudonocardiaceae					
					-- Amycolatopsis					-4.628233e-02
					`-- Amycolatopsis mediterranei				-5.537711e-05
				`-- Saccharopolyspora					-5.529852e-02
				`-- Saccharopolyspora erythraea					-1.794951e-04
			|-- Segniliparaceae					1.439771e-02
				`-- Segniliparus					4.798019e-05
				`-- Segniliparus rotundus					5.628289e-07
			|-- Streptomycetaceae					-8.012754e-02
				`-- Streptomyces					-5.876193e-05
				`-- Streptomyces scabiei					-5.982292e-02
			`--						
			`-- Tropheryma						-8.037162e-03
			`-- Tropheryma whipplei						-6.699194e-05
		|-- Bifidobacteriales						
			`-- Bifidobacteriaceae						
			`-- Bifidobacterium						
			`-- Bifidobacterium adolescentis						-1.081100e-03
		`-- Coriobacteriales							6.015406e-02
		`-- Coriobacteriaceae							1.145845e-04
	`-- Bacteroidetes					6.893292e-03			
	|-- Bacteroidia							
		`-- Bacteroidales							
			-- Prevotellaceae							
			`-- Prevotella							
			`-- Prevotella ruminicola						-1.837295e-02
		`--							
		`-- Candidatus Azobacteroides		-7.792114e-02				-1.354073e-01
			`-- Candidatus Azobacteroides pseudotrichonymphae						-1.007540e-04
	|-- Cytophagia				2.373230e-02			
		`-- Cytophagales				5.118349e-06			
			-- Cytophagaceae							
			`-- Spirosoma						-2.029576e-02
			`-- Spirosoma linguale						-3.593903e-04
		`-- Flammeovirgaceae							1.846172e-02
		`-- Marivirga							9.356407e-05
		`-- Marivirga tractuosa							1.089938e-07
	|-- Flavobacteria			6.597662e-02	3.598802e-02			
		`-- Flavobacteriales			2.848417e-05	1.292338e-05			
			-- Flavobacteriaceae							
				-- Flavobacterium							
				`-- Flavobacterium psychrophilum						
				-- Riemerella							6.527646e-03
				`-- Riemerella anatipestifer							3.285054e-07
			`-- Zunongwangia			9.813511e-03			
		`--							
		`-- Candidatus Sulcia			6.892260e-02				9.248245e-02
			`-- Candidatus Sulcia muelleri			1.196990e-04			
	`-- Sphingobacteria								-2.111405e-02
	`-- Sphingobacteriales								-6.646259e-05
	`-- Rhodothermaceae								
	`-- Salinibacter				-1.810959e-03				-1.562156e-01
	`-- Salinibacter ruber				-3.680658e-05				-1.425613e-03
|-- Cyanobacteria								
	`--								
		--								
		`--								
		`-- Acaryochloris			1.232690e-01				
		`-- Acaryochloris marina			1.546638e-04				
	|-- Chroococcales							
		`--							
			-- Cyanothece							
				-- Cyanothece sp. ATCC 51142						-6.097249e-02
				-- Cyanothece sp. PCC 7424						-1.200149e-02
			`-- Cyanothece sp. PCC 8801						1.510874e-02
		|-- Synechococcus						-2.562986e-02
			`-- Synechococcus sp. CC9311						-4.431117e-03
		`--							
		`-- Cyanobacterium UCYN-A			9.995647e-03				
	|-- Nostocales							
		`-- Nostocaceae							
			-- Anabaena							
				-- Anabaena azollae			-2.486573e-01	-1.317102e-01	-1.140282e-01	-3.488580e-01
			`-- Anabaena variabilis						
		`-- Nostoc							-2.903920e-02
	|-- Oscillatoriales			-1.758581e-01				-7.477761e-02
		`--							
		`-- Trichodesmium			-9.547240e-05				-7.845889e-06
		`-- Trichodesmium erythraeum			-9.944768e-05				-2.409806e-04
	`-- Prochlorales				4.356852e-02	1.784556e-02			3.063430e-03
	`-- Prochlorococcaceae					3.485743e-04			2.690010e-06
	`-- Prochlorococcus				1.524564e-04	8.515360e-06			1.475711e-05



|-- Betaproteobacteria						
| |-- Burkholderiales						
| | |-- Burkholderiaceae						
| | | |-- Burkholderia						
| | | | |-- Burkholderia ambifaria						4.965392e-02
| | | | |-- Burkholderia cenocepacia						9.953303e-04
| | | | |-- Burkholderia glumae			-5.088892e-02			-3.308498e-02
| | | | |-- Burkholderia mallei						-2.825884e-02
| | | | |-- Burkholderia phytofirmans						1.215842e-02
| | | | |-- Burkholderia pseudomallei						-3.993024e-02
| | | | |-- Burkholderia rhizoxinica			-4.104001e-03			-1.644402e-01
| | | | |-- Burkholderia sp. CCGE1002						2.111122e-02
| | | | |-- Burkholderia sp. CCGE1003						2.247689e-02
| | | | \`-- Burkholderia thailandensis						-8.637331e-03
| | | \`-- Ralstonia						1.002058e-02
| | | \`-- Ralstonia pickettii						4.128443e-02
| | |-- Comamonadaceae						
| | | |-- Alicycliphilus						2.659577e-02
| | | |-- Polaromonas			-2.878594e-02			-2.708253e-02
| | | | \`-- Polaromonas naphthalenivorans			-2.988079e-02			-2.774337e-02
| | | \`-- Verminephrobacter						-1.316877e-02
| | | \`-- Verminephrobacter eiseniae						-1.265057e-04
| | \`-- Oxalobacteraceae						
| | \`-- Candidatus Zinderia			-2.131300e-01	-1.191488e-01	-7.977046e-02	-3.136030e-01
| | \`-- Candidatus Zinderia insecticola			-5.334338e-05	-2.185004e-04	-1.988803e-04	-7.282768e-04
| |-- Gallionellales						3.999959e-02
| | \`-- Gallionellaceae						8.456717e-05
| |-- Neisseriales						
| | \`-- Neisseriaceae						
| | \`-- Neisseria						
| | |-- Neisseria meningitidis			-9.826837e-03			
| | \`-- Neisseria gonorrhoeae				-6.534861e-03		-9.794726e-02
| |-- Methylophilales			1.827182e-04			1.744261e-02
| | \`-- Methylophilaceae						5.115326e-05
| | \`-- Methylotenera						8.924738e-03
| \`--						
| \`--						
| \`-- Candidatus Accumulibacter						-4.200990e-03
| \`-- Candidatus Accumulibacter phosphatis						-1.787884e-05
|-- Gammaproteobacteria						-5.382801e-03
| |-- Alteromonadales						
| | |-- Alteromonadaceae						
| | | \`-- Alteromonas			8.173098e-02			
| | | \`-- Alteromonas macleodii			1.069440e-04			
| | |-- Pseudoalteromonadaceae			1.367971e-02			1.409535e-02
| | | \`-- Pseudoalteromonas			6.703904e-06			6.592843e-05
| | \`-- Shewanellaceae						
| | \`-- Shewanella						
| | |-- Shewanella baltica						9.999962e-03
| | |-- Shewanella woodyi				-3.914312e-01	-3.245567e-01	
| | \`-- Shewanella sp. MR-7			-5.191838e-03			
| |-- Chromatiales						
| | |-- Chromatiaceae						
| | | \`-- Nitrosococcus						
| | | \`-- Nitrosococcus oceani						-2.250092e-03
| | \`-- Ectothiorhodospiraceae						4.971180e-02
| | \`-- Thioalkalivibrio						1.961641e-02
| |-- Enterobacteriales						-1.768873e-02
| | \`-- Enterobacteriaceae						-2.646109e-06
| | |-- Buchnera			1.653598e-02	4.034737e-02		8.866412e-02
| | | \`-- Buchnera aphidicola			8.222839e-06	2.784827e-05		
| | |-- Cronobacter						-4.011128e-02
| | | |-- Cronobacter sakazakii				5.294979e-02		
| | | \`-- Cronobacter turicensis						-1.893411e-02
| | |-- Edwardsiella			7.245479e-02			
| | |-- Enterobacter						
| | | \`-- Enterobacter cloacae			1.186288e-02			9.075749e-03
| | |-- Erwinia						
| | | |-- Erwinia amylovora						-5.762296e-02
| | | \`-- Erwinia pyrifoliae			-1.775894e-02	-5.111413e-02		-4.958100e-03
| | |-- Escherichia				-7.204872e-03	-1.167991e-02		-4.598495e-02
| | | \`-- Escherichia coli			-3.289986e-02	-9.126601e-04		
| | |-- Pantoea						
| | | |-- Pantoea ananatis						-5.746220e-02
| | | \`-- Pantoea sp. At-9b						1.045491e-01
| | |-- Photorhabdus						
| | | \`-- Photorhabdus luminescens					-2.806162e-01	-2.019716e-01	
| | |-- Proteus						6.451190e-03
| | |-- Salmonella			4.103872e-03			-1.659207e-03
| | | \`-- Salmonella enterica						-3.003756e-05
| | |-- Sodalis			-6.680653e-02	-2.610751e-01	-2.281354e-01	-1.828575e-01	-3.829020e-01
| | | \`-- Sodalis glossinidius			-4.666137e-04	-6.919945e-04	-8.413410e-04	-3.825242e-04	-6.079117e-04
| | |-- Shigella				-5.526016e-03	-4.419651e-02	-3.999940e-02	-1.037834e-01
| | | |-- Shigella boydii							
| | | |-- Shigella dysenteriae				-5.736233e-02	-1.769379e-03		
| | | \`-- Shigella sonnei				-9.094229e-02			-6.755381e-03
| | |-- Wigglesworthia							2.605822e-02
| | |-- Yersinia							
| | | \`-- Yersinia pestis				-2.088408e-02			
| | |-- Candidatus Blochmannia							5.592574e-02
| | |-- Candidatus Riesia				6.068940e-02			
| | | \`-- Candidatus Riesia pediculicola				3.809398e-05			
| | \`-- Citrobacter					9.842755e-04		
| | \`-- Citrobacter koseri				-3.952175e-02	2.904362e-03		
| |-- Pasteurellales			8.939955e-02				
| | \`-- Pasteurellaceae			5.435622e-05				
| | |-- Actinobacillus							
| | | \`-- Actinobacillus pleuropneumoniae				-3.821200e-02			-2.408705e-02
| | \`-- Aggregatibacter			8.879968e-02				3.524730e-02
| | \`-- Aggregatibacter actinomycetemcomitans				2.664342e-03			1.608978e-02
| |-- Pseudomonadales							-4.618728e-02
| | |-- Moraxellaceae							
| | | |-- Acinetobacter							
| | | \`-- Acinetobacter baumannii				-1.125266e-02	-2.930301e-02	-1.506712e-02	-7.211917e-02
| | | \`-- Psychrobacter							
| | | \`-- Psychrobacter sp. PRwf-1				-9.492501e-02			
| | \`-- Pseudomonadaceae							
| | \`-- Pseudomonas							
| | \`-- Pseudomonas savastanoi							-2.277981e-02
| |-- Thiotrichales							
| | \`-- Francisellaceae							
| | \`-- Francisella							
| | \`-- Francisella tularensis							-5.562229e-02
| |-- Vibrionales							
| | \`-- Vibrionaceae							

```
|   |   |   |       |-- Aliivibrio                                  |-------------|-------------|-------------|-------------|-------------|
|   |   |   |       |   |-- Aliivibrio fischeri                     |-------------| 5.850647e-02|-------------|-------------| 6.213608e-02|
|   |   |   |       |   `-- Aliivibrio salmonicida                  |-------------|-7.829542e-04|-------------|-------------|-3.193901e-02|
|   |   |   |       |-- Photobacterium                              |-------------|-------------|-------------|-------------|-5.184353e-02|
|   |   |   |       |   `-- Photobacterium profundum                |-------------|-------------|-------------|-------------|-8.417983e-04|
|   |   |   |       `-- Vibrio                                      |-------------|-------------|-------------|-------------|-------------|
|   |   |   |           |-- Vibrio cholerae                         |-------------|-------------|-------------|-------------|-3.486180e-02|
|   |   |   |           |-- Vibrio harveyi                          |-------------|-------------| 9.105963e-02|-------------|-------------|
|   |   |   |           `-- Vibrio vulnificus                       |-------------|-------------|-------------|-------------|-1.017048e-02|
|   |   |   |-- Xanthomonadales                                     |-------------|-------------|-5.763670e-02|-5.097082e-02|-------------|
|   |   |   |   `-- Xanthomonadaceae                                |-------------|-------------|-1.591782e-03|-1.939096e-04|-------------|
|   |   |   |       |-- Pseudoxanthomonas                           |-------------|-------------|-------------|-------------| 1.767922e-03|
|   |   |   |       |-- Xanthomonas                                 |-4.132276e-02|-1.189451e-02|-------------|-------------|-1.099975e-01|
|   |   |   |       |   |-- Xanthomonas axonopodis                  |-------------|-------------|-------------|-------------|-9.111066e-02|
|   |   |   |       |   `-- Xanthomonas oryzae                      |-------------|-------------|-------------|-------------|-1.018949e-01|
|   |   |   |       `-- Xylella                                     |-------------|-------------|-3.831987e-02|-1.752824e-02|-------------|
|   |   |   |           `-- Xylella fastidiosa                      |-------------|-------------|-1.101869e-07|-2.231971e-04|-------------|
|   |   |   `--                                                     |-------------|-------------|-------------|-------------|-------------|
|   |   |       `--                                                 |-------------|-------------|-------------|-------------|-------------|
|   |   |           `-- Candidatus Carsonella                       |-------------| 4.285463e-02|-------------|-------------| 3.131953e-02|
|   |   |               `-- Candidatus Carsonella ruddii            |-------------| 8.276704e-06|-------------|-------------|-------------|
|   |   |-- Epsilonproteobacteria                                   | 3.368466e-02| 5.331978e-02| 4.548317e-02| 4.134447e-02| 8.776522e-02|
|   |   |   |-- Campylobacterales                                   |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |-- Campylobacteraceae                              |-------------| 4.944910e-02|-------------|-------------|-------------|
|   |   |   |   |   |-- Arcobacter                                  |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |   |   `-- Arcobacter butzleri                     |-------------|-------------|-------------|-3.716494e-01|-------------|
|   |   |   |   |   `-- Campylobacter                               |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |       `-- Campylobacter hominis                   |-------------| 2.788010e-02|-------------|-------------|-------------|
|   |   |   |   `-- Helicobacteraceae                               |-------------|-------------|-------------|-------------|-------------|
|   |   |   |       `-- Sulfuricurvum                               |-------------|-------------|-------------|-------------| 2.186217e-02|
|   |   |   |           `-- Sulfuricurvum kujiense                  |-------------|-------------|-------------|-------------|-------------|
|   |   |   `-- Legionellales                                       |-------------|-------------|-------------|-------------|-------------|
|   |   |       `-- Coxiellaceae                                    |-------------|-3.891974e-02|-1.471500e-02|-1.275228e-02|-1.533101e-01|
|   |   |           `-- Coxiella                                    |-------------|-3.781573e-04|-1.739126e-06|-6.458707e-07|-------------|
|   |   |               `-- Coxiella burnetii                       |-------------|-9.051066e-05|-6.495941e-05|-2.708368e-05|-3.518029e-03|
|   |   `-- Deltaproteobacteria                                     |-------------|-1.353730e-02|-------------|-------------|-------------|
|   |       |-- Desulfobacterales                                   |-------------|-------------|-------------|-------------|-------------|
|   |       |   `-- Desulfobacteraceae                              |-------------|-------------|-------------|-------------| 3.001447e-02|
|   |       |       `-- Desulfatibacillum                           |-------------|-------------|-3.756557e-01|-3.034182e-01|-------------|
|   |       |           `-- Desulfatibacillum alkenivorans          |-------------|-------------|-3.707260e-04|-1.246228e-03|-------------|
|   |       |-- Desulfovibrionales                                  |-------------|-------------|-------------|-------------|-------------|
|   |       |   `-- Desulfovibrionaceae                             |-------------|-------------|-------------|-------------|-------------|
|   |       |       `-- Desulfovibrio                               |-------------|-------------|-------------|-2.991914e-03|-------------|
|   |       |           |-- Desulfovibrio magneticus                |-------------|-------------|-------------|-------------|-3.326707e-02|
|   |       |           `-- Desulfovibrio vulgaris                  |-------------|-------------|-------------|-7.288790e-02|-------------|
|   |       |-- Desulfuromonadales                                  |-------------|-------------|-------------|-------------|-------------|
|   |       |   |-- Geobacteraceae                                  |-------------|-------------|-------------|-------------|-------------|
|   |       |   |   `-- Geobacter                                   |-------------|-------------|-------------|-------------|-------------|
|   |       |   |       `-- Geobacter lovleyi                       |-------------|-------------|-------------|-------------| 3.451219e-02|
|   |       |   `-- Pelobacteraceae                                 |-------------|-------------|-------------|-------------|-------------|
|   |       |       `-- Pelobacter                                  |-------------|-------------|-------------|-------------|-------------|
|   |       |           `-- Pelobacter propionicus                  |-------------|-------------|-------------|-------------|-6.348945e-02|
|   |       |-- Myxococcales                                        |-------------|-5.738846e-02|-------------|-------------|-------------|
|   |       |   |-- Myxococcaceae                                   |-------------|-------------|-------------|-------------|-------------|
|   |       |   |   `-- Myxococcus                                  |-------------|-------------|-------------|-------------|-2.257533e-01|
|   |       |   |       `-- Myxococcus xanthus                      |-------------|-------------|-------------|-------------|-3.055500e-03|
|   |       |   `-- Polyangiaceae                                   |-------------|-------------|-------------|-------------|-2.411446e-01|
|   |       |       `-- Sorangium                                   |-------------|-------------|-------------|-------------|-2.737126e-04|
|   |       |           `-- Sorangium cellulosum                    |-------------|-------------|-------------|-------------|-3.026196e-03|
|   |       `-- Syntrophobacterales                                 |-------------|-------------|-------------|-------------|-7.612898e-03|
|   |-- Firmicutes                                                  |-------------|-------------|-------------|-------------| 1.750367e-02|
|   |   |-- Bacilli                                                 |-------------|-------------|-------------|-------------|-------------|
|   |   |   |-- Bacillales                                          |-------------|-------------|-1.220667e-03|-------------|-------------|
|   |   |   |   |-- Alicyclobacillaceae                             |-------------|-2.700911e-02|-5.544236e-02|-------------|-------------|
|   |   |   |   |-- Bacillaceae                                     |-------------|-2.411869e-02|-3.610897e-02|-------------|-1.556876e-02|
|   |   |   |   |   |-- Bacillus                                    |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |   |   |-- Bacillus licheniformis                  |-------------|-------------|-------------|-2.449997e-02|-------------|
|   |   |   |   |   |   |-- Bacillus megaterium                     |-------------|-------------|-------------|-9.783724e-02|-------------|
|   |   |   |   |   |   |-- Bacillus pseudofirmus                   |-------------|-------------|-------------|-------------| 1.554171e-03|
|   |   |   |   |   |   |-- Bacillus subtilis                       |-------------|-------------|-------------|-2.863488e-02|-------------|
|   |   |   |   |   |   `-- Bacillus thuringiensis                  |-------------|-------------|-------------|-------------|-5.523809e-02|
|   |   |   |   |   |-- Anoxybacillus                               |-------------|-------------|-------------|-------------|-1.131978e-01|
|   |   |   |   |   |   `-- Anoxybacillus flavithermus              |-------------|-------------|-------------|-------------|-6.532787e-04|
|   |   |   |   |   |-- Geobacillus                                 |-------------|-6.382519e-02|-3.164481e-02|-------------|-------------|
|   |   |   |   |   |   |-- Geobacillus kaustophilus                |-------------|-------------|-------------|-------------|-2.814032e-03|
|   |   |   |   |   |   |-- Geobacillus thermodenitrificans         |-------------|-------------|-------------|-------------|-6.995346e-02|
|   |   |   |   |   |   `-- Geobacillus sp. WCH70                   |-------------|-3.600912e-02|-------------|-------------|-------------|
|   |   |   |   |   `-- Lysinibacillus                              |-------------| 1.708490e-01|-------------|-------------|-2.248163e-02|
|   |   |   |   |       `-- Lysinibacillus sphaericus               |-------------| 3.096908e-04|-------------|-------------|-9.173783e-05|
|   |   |   |   |-- Listeriaceae                                    | 7.502670e-03|-------------|-------------|-------------| 3.341008e-02|
|   |   |   |   |   `-- Listeria                                    | 4.326177e-06|-------------|-------------|-------------| 1.532926e-05|
|   |   |   |   |       `-- Listeria innocua                        |-------------|-------------|-------------|-6.485660e-02|-------------|
|   |   |   |   |-- Paenibacillaceae                                |-------------|-1.817487e-03|-------------|-------------|-------------|
|   |   |   |   |-- Staphylococcaceae                               |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |   |-- Macrococcus                                 |-------------|-1.079243e-01|-8.924683e-02|-5.757300e-02|-1.371328e-01|
|   |   |   |   |   |   `-- Macrococcus caseolyticus                |-------------|-1.136328e-03|-9.819945e-06|-7.754847e-04|-2.884859e-04|
|   |   |   |   |   `-- Staphylococcus                              |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |       |-- Staphylococcus aureus                   |-------------| 1.295175e-02|-------------| 1.770880e-02| 7.645958e-03|
|   |   |   |   |       |-- Staphylococcus epidermidis              |-------------|-------------|-------------|-------------|-3.887997e-03|
|   |   |   |   |       `-- Staphylococcus haemolyticus             |-------------|-------------|-------------|-------------| 8.241154e-02|
|   |   |   |   `--                                                 |-------------|-------------|-------------|-------------|-------------|
|   |   |   |       `-- Exiguobacterium                             |-------------|-------------|-------------|-------------| 9.204457e-02|
|   |   |   |           `-- Exiguobacterium sibiricum               |-------------|-------------|-------------|-------------| 3.272285e-02|
|   |   |   `-- Lactobacillales                                     | 1.137273e-02| 1.041380e-02|-------------| 5.191705e-03|-------------|
|   |   |       |-- Lactobacillaceae                                |-------------|-------------|-------------|-------------|-------------|
|   |   |       |   `-- Lactobacillus                               |-------------|-------------|-------------|-------------|-------------|
|   |   |       |       |-- Lactobacillus brevis                    |-------------|-2.664710e-02|-------------|-------------|-------------|
|   |   |       |       |-- Lactobacillus helveticus                |-------------|-------------|-------------|-------------|-4.148728e-03|
|   |   |       |       `-- Lactobacillus johnsonii                 |-------------|-------------|-------------|-------------| 2.576702e-02|
|   |   |       |-- Leuconostocaceae                                |-------------| 2.390437e-02|-------------|-------------|-------------|
|   |   |       |   `-- Leuconostoc                                 |-------------| 8.040167e-03|-------------|-------------| 6.062086e-03|
|   |   |       |       `-- Leuconostoc mesenteroides               |-------------|-------------|-------------|-------------|-------------|
|   |   |       `-- Streptococcaceae                                |-------------|-------------|-------------|-------------|-------------|
|   |   |           |-- Lactococcus                                 |-------------|-5.161360e-02|-------------|-------------|-2.669661e-02|
|   |   |           |   `-- Lactococcus lactis                      |-------------|-3.527781e-04|-------------|-------------|-2.737881e-04|
|   |   |           `-- Streptococcus                               | 6.371984e-02| 1.259744e-02| 1.291077e-02|-------------|-------------|
|   |   |               |-- Streptococcus pneumoniae                |-------------| 2.155931e-02|-------------|-------------|-------------|
|   |   |               |-- Streptococcus suis                      |-------------|-------------|-------------|-------------|-6.458698e-03|
|   |   |               `-- Streptococcus thermophilus              |-------------|-6.104892e-02|-------------|-------------|-------------|
|   |   |-- Clostridia                                              | 3.160940e-02|-------------|-------------|-------------| 2.653723e-02|
|   |   |   |-- Clostridiales                                       |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |-- Clostridiaceae                                  |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |   |-- Alkaliphilus                                |-------------|-------------|-------------|-4.675426e-01|-------------|
```

```
|   |   |   |   |   `-- Clostridium                                      |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |       |-- Clostridium botulinum                        |-------------| 3.323043e-02|-------------|-------------|-------------|
|   |   |   |   |       |-- Clostridium kluyveri                         |-------------|-------------|-------------|-------------|-1.188038e-02|
|   |   |   |   |       `-- Clostridium tetani                           |-------------|-1.630349e-02|-------------|-------------|-1.494068e-01|
|   |   |   |   |-- Eubacteriaceae                                       |-------------| 3.484944e-02| 8.384171e-02|-------------|-------------|
|   |   |   |   |   `-- Eubacterium                                      |-------------| 6.525040e-05| 5.579187e-06|-------------|-------------|
|   |   |   |   |       `-- Eubacterium eligens                          |-------------| 6.008522e-03| 9.019429e-03|-------------|-------------|
|   |   |   |   |-- Clostridiales Family XI. Incertae Sedis              |-------------|-------------|-------------|-------------| 3.197574e-02|
|   |   |   |   |-- Clostridiales Family XVIII. Incertae Sedis           |-------------|-------------|-------------|-------------|-4.939135e-02|
|   |   |   |   |   `-- Symbiobacterium                                  |-------------|-------------|-------------|-------------|-2.496554e-06|
|   |   |   |   |       `-- Symbiobacterium thermophilum                 |-------------|-------------|-------------|-------------|-7.133964e-05|
|   |   |   |   |-- Peptococcaceae                                       |-------------|-3.386838e-02|-------------|-4.878099e-02|-------------|
|   |   |   |   |   |-- Desulfitobacterium                               |-------------|-------------|-------------|-------------|-5.292300e-02|
|   |   |   |   |   |   `-- Desulfitobacterium hafniense                 |-------------|-------------|-------------|-------------|-3.500918e-04|
|   |   |   |   |   |-- Pelotomaculum                                    |-------------|-------------|-------------|-2.508862e-01|-------------|
|   |   |   |   |   |   `-- Pelotomaculum thermopropionicum              |-------------|-------------|-------------|-2.885571e-04|-------------|
|   |   |   |   |   |-- Thermincola                                      |-------------|-------------|-------------|-2.635677e-01|-------------|
|   |   |   |   |   |   `-- Thermincola potens                           |-------------|-------------|-------------|-1.131690e-03|-------------|
|   |   |   |   |   `-- Candidatus Desulforudis                          |-------------|-------------|-------------|-2.270982e-01|-------------|
|   |   |   |   |       `-- Candidatus Desulforudis audaxviator          |-------------|-------------|-------------|-3.290372e-04|-------------|
|   |   |   |   |-- Ruminococcaceae                                      |-------------|-3.008062e-02|-------------|-------------|-------------|
|   |   |   |   `-- Syntrophomonadaceae                                  |-------------|-------------|-------------|-------------|-------------|
|   |   |   |       `-- Syntrophomonas                                   |-------------|-------------|-------------|-3.102357e-01|-------------|
|   |   |   |           `-- Syntrophomonas wolfei                        |-------------|-------------|-------------|-1.243938e-03|-------------|
|   |   |   |-- Halanaerobiales                                          |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   `-- Halobacteroidaceae                                   |-------------|-------------|-------------|-3.163949e-01|-------------|
|   |   |   |       `-- Acetohalobium                                    |-------------|-------------|-------------|-6.609357e-06|-------------|
|   |   |   |           `-- Acetohalobium arabaticum                     |-------------|-------------|-------------|-6.304524e-04|-------------|
|   |   |   `-- Thermoanaerobacterales                                   |-------------|-------------|-------------|-------------|-------------|
|   |   |       |-- Thermoanaerobacteraceae                              |-------------|-------------|-------------|-------------|-------------|
|   |   |       |   |-- Moorella                                         |-------------|-------------|-------------|-2.877254e-01|-------------|
|   |   |       |   |   `-- Moorella thermoacetica                       |-------------|-------------|-------------|-1.178556e-03|-------------|
|   |   |       |   `-- Thermoanaerobacter                               |-------------|-------------|-------------|-------------|-------------|
|   |   |       |       `-- Thermoanaerobacter mathranii                 |-------------|-------------|-------------|-3.180552e-01|-------------|
|   |   |       `-- Thermoanaerobacterales Family III. Incertae Sedis    |-------------|-4.194183e-03|-------------|-------------| 1.134667e-02|
|   |   `-- Negativicutes                                                |-------------|-------------|-------------|-------------| 3.843326e-02|
|   |       `-- Selenomonadales                                          |-------------|-------------|-------------|-------------| 3.803060e-06|
|   |-- Chlorobi                                                         |-------------|-------------|-------------|-------------|-------------|
|   |   `-- Chlorobia                                                    |-------------|-------------|-------------|-------------|-------------|
|   |       `-- Chlorobiales                                             |-------------|-------------|-------------|-------------|-------------|
|   |           `-- Chlorobiaceae                                        |-------------|-------------|-------------|-------------|-------------|
|   |               `-- Chlorobium                                       |-------------|-------------|-------------|-------------|-------------|
|   |                   `-- Chlorobium chlorochromatii                   |-------------|-------------|-3.935665e-01|-3.064457e-01|-------------|
|   |-- Chloroflexi                                                      |-------------|-7.708783e-03|-------------|-------------|-------------|
|   |   |-- Chloroflexi                                                  |-------------|-------------|-9.718679e-03|-------------|-------------|
|   |   |   `-- Chloroflexales                                           |-1.072941e-01|-------------|-------------|-------------|-------------|
|   |   |       `-- Chloroflexaceae                                      |-------------|-------------|-------------|-------------|-------------|
|   |   |           |-- Chloroflexus                                     |-1.921534e-01|-------------|-------------|-------------|-------------|
|   |   |           `-- Roseiflexus                                      |-------------|-------------|-------------|-------------|-1.372706e-02|
|   |   |-- Dehalococcoidetes                                            | 1.582225e-02| 1.737766e-02|-------------|-------------| 6.466606e-02|
|   |   |   `--                                                          |-------------|-------------|-------------|-------------|-------------|
|   |   |       `--                                                      |-------------|-------------|-------------|-------------|-------------|
|   |   |           `-- Dehalococcoides                                  |-------------| 7.603946e-04|-------------|-------------|-------------|
|   |   `-- Thermomicrobia                                               |-------------|-------------|-------------|-------------| 1.478051e-02|
|   |       |-- Sphaerobacterales                                        |-------------|-------------|-------------|-------------| 5.254531e-02|
|   |       |   `-- Sphaerobacteraceae                                   |-------------|-------------|-------------|-------------| 1.679418e-04|
|   |       |       `-- Sphaerobacter                                    |-------------|-------------|-------------|-------------| 1.133596e-04|
|   |       |           `-- Sphaerobacter thermophilus                   |-------------|-------------|-------------|-------------| 1.841835e-06|
|   |       `-- Thermomicrobiales                                        |-------------| 2.488133e-02|-------------|-------------|-------------|
|   |-- Chlamydiae                                                       |-------------|-------------|-------------|-------------|-------------|
|   |   `-- Chlamydiae                                                   |-------------|-------------|-------------|-------------|-------------|
|   |       `-- Chlamydiales                                             |-------------|-------------|-------------|-------------|-------------|
|   |           `-- Chlamydiaceae                                        |-------------|-------------|-------------|-------------| 6.362852e-03|
|   |               |-- Chlamydia                                        |-------------|-------------|-------------|-------------|-------------|
|   |               |   `-- Chlamydia trachomatis                        |-------------| 1.953302e-02|-------------|-------------| 6.006223e-02|
|   |               `-- Chlamydophila                                    |-------------|-------------|-------------|-------------|-------------|
|   |                   `-- Chlamydophila felis                          |-------------|-------------|-------------|-------------|-3.935655e-02|
|   |-- Deferribacteres                                                  |-------------|-------------|-------------|-------------| 5.955722e-02|
|   |   `-- Deferribacteres                                              |-------------|-------------|-------------|-------------| 1.866701e-04|
|   |       `-- Deferribacterales                                        |-------------|-------------|-------------|-------------| 1.238492e-03|
|   |           `-- Deferribacteraceae                                   |-------------|-------------|-------------|-------------| 2.682359e-05|
|   |-- Spirochaetes                                                     |-------------|-------------|-------------|-------------|-------------|
|   |   `-- Spirochaetes                                                 |-------------|-------------|-------------|-------------|-------------|
|   |       `-- Spirochaetales                                           |-------------|-------------|-------------|-------------|-------------|
|   |           |-- Brachyspiraceae                                      |-------------| 1.077062e-01|-------------|-------------|-------------|
|   |           |-- Leptospiraceae                                       |-------------|-------------|-------------|-------------|-------------|
|   |           |   `-- Leptospira                                       |-------------|-------------|-------------|-------------|-------------|
|   |           |       `-- Leptospira borgpetersenii                    |-------------|-8.740157e-03|-------------|-------------|-------------|
|   |           `-- Spirochaetaceae                                      |-------------|-------------|-------------|-------------|-------------|
|   |               `-- Borrelia                                         |-9.516564e-02|-------------|-------------|-------------|-------------|
|   |                   |-- Borrelia afzelii                             |-------------|-------------| 3.547221e-03|-------------|-------------|
|   |                   |-- Borrelia bavariensis                         |-------------|-------------|-------------|-------------|-6.921856e-02|
|   |                   |-- Borrelia duttonii                            |-------------|-9.313604e-02|-1.182413e-01|-1.212633e-01|-1.548067e-01|
|   |                   `-- Borrelia recurrentis                         |-2.390453e-02|-2.229828e-01|-2.146843e-01|-2.096367e-01|-2.618717e-01|
|   |-- Synergistetes                                                    |-------------|-------------|-------------|-------------| 4.847815e-02|
|   |   `-- Synergistia                                                  |-------------|-------------|-------------|-------------| 2.711175e-05|
|   |       `-- Synergistales                                            |-------------|-------------|-------------|-------------| 3.245959e-06|
|   |           `-- Synergistaceae                                       |-------------|-------------|-------------|-------------| 1.243171e-06|
|   |-- Tenericutes                                                      |-------------|-------------|-------------|-------------|-------------|
|   |   `-- Mollicutes                                                   |-------------|-------------|-------------|-------------|-------------|
|   |       |-- Mycoplasmatales                                          |-3.716608e-01|-------------|-------------|-------------|-------------|
|   |       |   `-- Mycoplasmataceae                                     |-7.385837e-06|-------------|-------------|-------------|-------------|
|   |       |       |-- Mycoplasma                                       |-------------| 2.213015e-02|-------------|-------------| 2.205010e-02|
|   |       |       |   |-- Mycoplasma conjunctivae                      |-------------| 8.920143e-03|-------------|-------------|-------------|
|   |       |       |   |-- Mycoplasma hyopneumoniae                     |-------------|-6.675468e-02|-------------|-------------|-8.644123e-02|
|   |       |       |   `-- Mycoplasma synoviae                          |-------------|-1.518909e-02|-------------|-------------|-1.127758e-01|
|   |       |       `-- Ureaplasma                                       |-------------|-4.436889e-01|-4.216600e-01|-3.908682e-01|-5.495712e-01|
|   |       |-- Entomoplasmatales                                        |-------------|-3.772003e-01|-3.099263e-01|-2.277016e-01|-4.976261e-01|
|   |       |   `-- Entomoplasmataceae                                   |-------------|-9.391740e-05|-2.168236e-03|-4.966970e-05|-8.963129e-04|
|   |       |       `-- Mesoplasma                                       |-------------|-3.033416e-05|-1.741487e-04|-9.664716e-05|-4.035807e-04|
|   |       |           `-- Mesoplasma florum                            |-------------|-2.726959e-04|-3.102367e-04|-3.502865e-04|-1.309935e-03|
|   |       `-- Acholeplasmatales                                        |-------------|-------------|-------------|-------------|-------------|
|   |           `-- Acholeplasmataceae                                   |-------------|-------------|-------------|-------------|-------------|
|   |               `-- Candidatus Phytoplasma                           |-------------|-------------|-------------|-------------|-------------|
|   |                   `-- Aster yellows witches'-broom phytoplasma     | 1.170936e-01| 8.050883e-02| 3.437526e-02|-------------| 9.686432e-02|
|   |-- Thermotogae                                                      | 5.004699e-02|-------------|-------------|-------------| 3.299572e-02|
|   |   `-- Thermotogae                                                  | 5.913261e-05|-------------|-------------|-------------| 1.663027e-04|
|   |       `-- Thermotogales                                            | 6.197371e-04|-------------|-------------|-------------| 1.669198e-05|
|   |           `-- Thermotogaceae                                       | 3.718842e-05|-------------|-------------|-------------| 7.484175e-06|
|   |               `-- Thermotoga                                       |-------------|-------------|-------------|-------------|-------------|
|   |                   `-- Thermotoga neapolitana                       |-------------|-------------|-------------|-------------|-1.977506e-02|
|   |-- Verrucomicrobia                                                  |-------------|-------------|-------------|-------------|-------------|
```

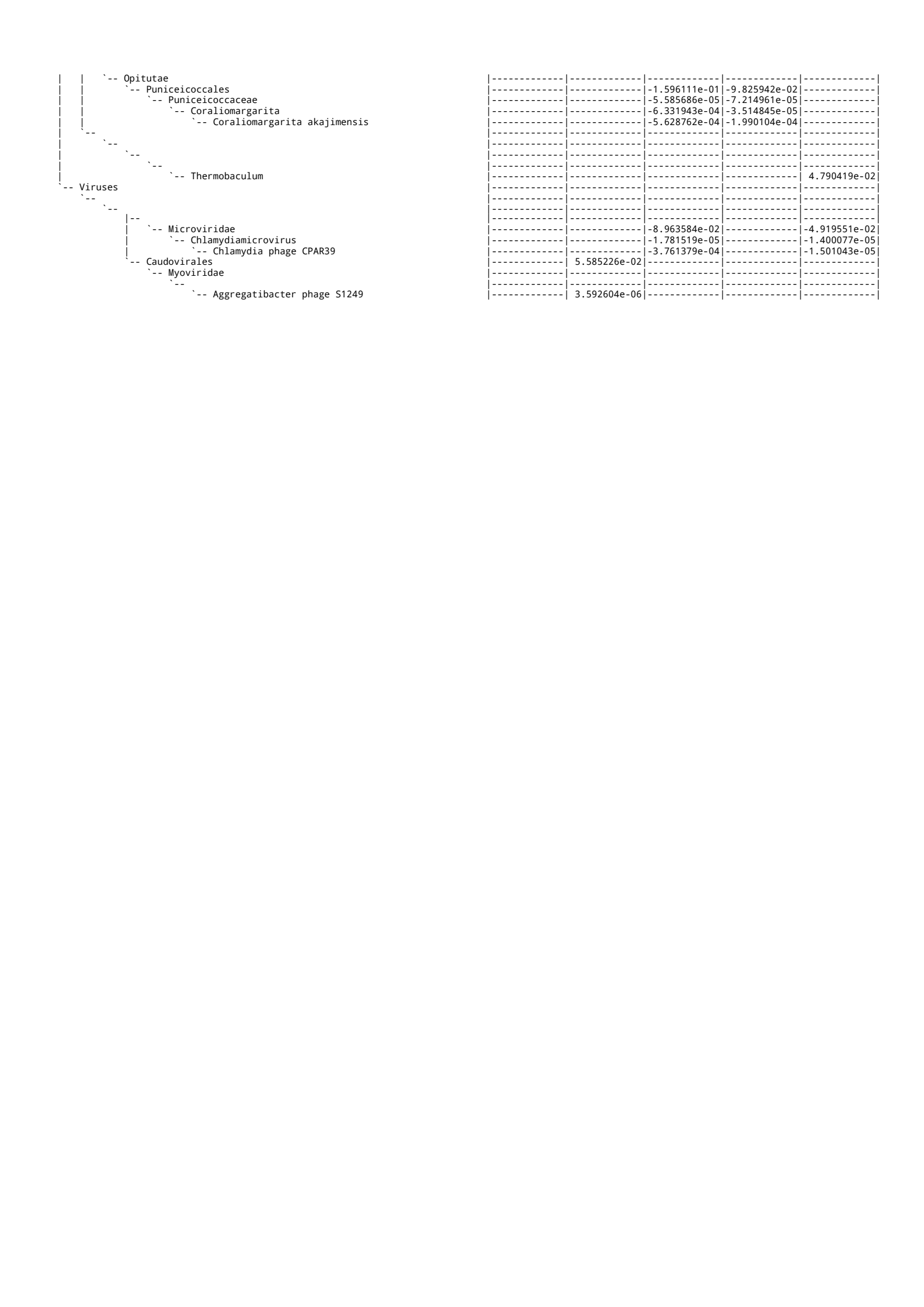
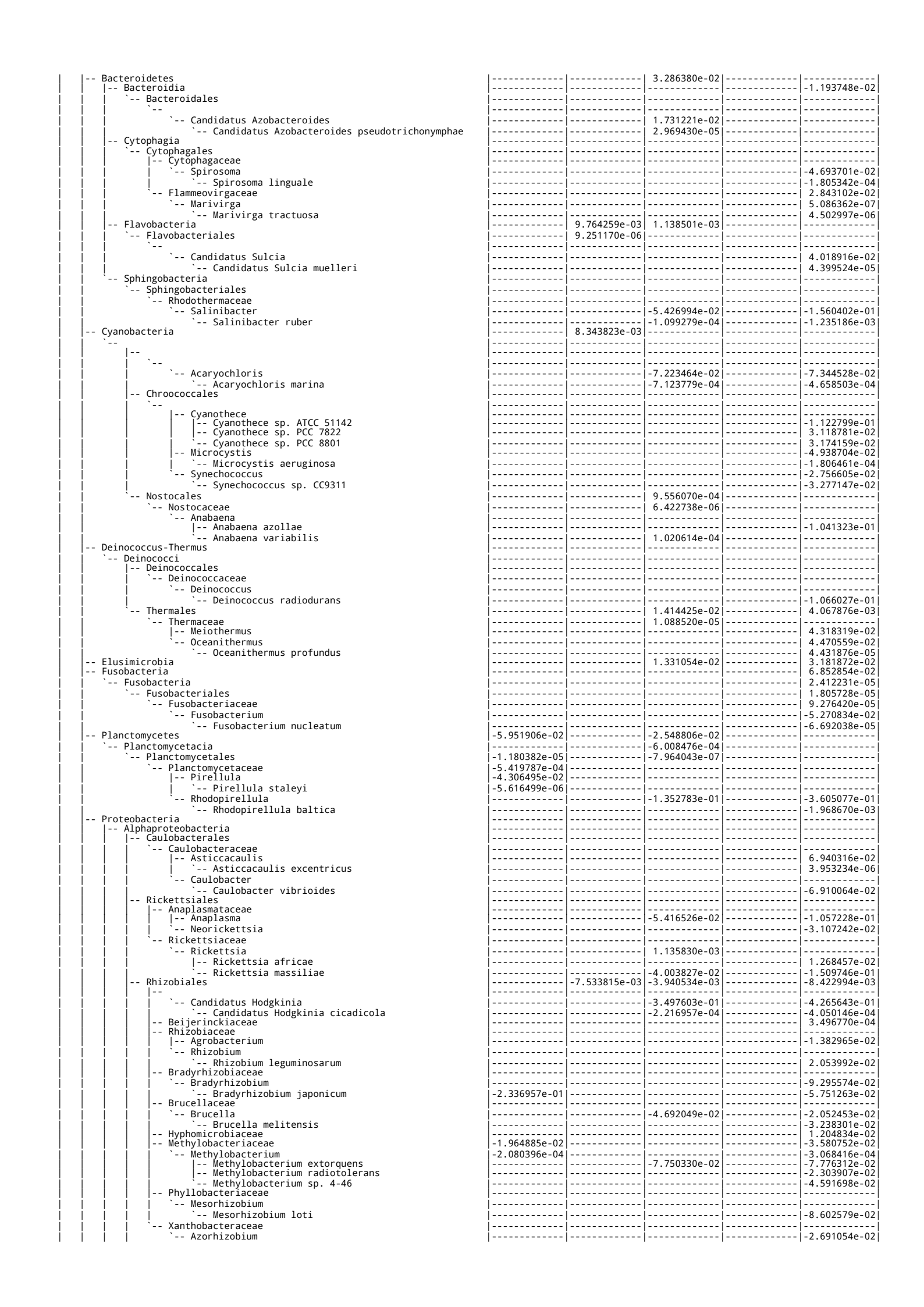

| | `-- Opitutae
| | `-- Puniceicoccales
| | `-- Puniceicoccaceae
| | `-- Coraliomargarita
| | `-- Coraliomargarita akajimensis
| `--
| `--
| `--
| `-- Thermobaculum
`-- Viruses
    `--
        `--
            |--
            | `-- Microviridae
            | `-- Chlamydiamicrovirus
            | `-- Chlamydia phage CPAR39
            `-- Caudovirales
                `-- Myoviridae
                    `--
                        `-- Aggregatibacter phage S1249
|-------------|-------------|-------------|-------------|-------------|
-------------	-------------	-1.596111e-01	-9.825942e-02	-------------
-------------	-------------	-5.585686e-05	-7.214961e-05	-------------
-------------	-------------	-6.331943e-04	-3.514845e-05	-------------
-------------	-------------	-5.628762e-04	-1.990104e-04	-------------
-------------	-------------	-------------	-------------	4.790419e-02
-------------	-------------	-8.963584e-02	-------------	-4.919551e-02
-------------	-------------	-1.781519e-05	-------------	-1.400077e-05
-------------	-------------	-3.761379e-04	-------------	-1.501043e-05
-------------	5.585226e-02	-------------	-------------	-------------
-------------	3.592604e-06	-------------	-------------	-------------

Table 4.4: Lasso coefficients for recall (5'+3')

The table shows the coefficients produced by performing the lasso on the recall model for exact matching gene start and stop. The coefficients for each program are for the model chosen by the 'one-standard-error' rule (see section 3.4.2).
See also figure 4.8 for more information about values of the shrinkage factors.

Coefficients	MED	Glimmmer	GeneMark.hmm	GeneMarkS	Prodigal
Intercept	5.855028e-01	7.837815e-01	8.128980e-01	8.253003e-01	8.203130e-01
GC	-6.627416e-02	-1.782160e-01	-1.721235e-01	-1.875599e-01	-9.365103e-02
Length (bp)	4.637104e-08	4.538357e-11	5.271760e-09	1.434605e-09	2.291151e-08
Taxonomy					
-- Archaea			1.441930e-03		
|-- Crenarchaeota					
| `-- Thermoprotei					
| `-- Desulfurococcales					
| `-- Desulfurococcaceae					
| `-- Staphylothermus					
| `-- Staphylothermus hellenicus				-2.949777e-02	
`-- Euryarchaeota					
|-- Halobacteria			3.525388e-02		
| `-- Halobacteriales			2.038598e-04		
| `-- Halobacteriaceae			1.791442e-05		
| |-- Halalkalicoccus					-1.185831e-01
| | `-- Halalkalicoccus jeotgali					-6.072651e-04
| |-- Haloarcula					-4.462351e-02
| | `-- Haloarcula marismortui					-2.688984e-04
| |-- Haloferax					-1.053753e-03
| |-- Haloterrigena			1.986965e-02		1.042920e-02
| | `-- Haloterrigena turkmenica			1.741355e-06		3.468007e-05
| `-- Natrialba					4.432276e-02
| `-- Natrialba magadii					7.290351e-05
|-- Methanobacteria					
| `-- Methanobacteriales					
| `-- Methanothermaceae					
| `-- Methanothermobacter					-3.840771e-02
|-- Methanococci					5.543408e-03
| `-- Methanococcales					4.352358e-05
| `-- Methanocaldococcaceae					
| `-- Methanocaldococcus					
| |-- Methanocaldococcus jannaschii					-3.396182e-02
| `-- Methanocaldococcus sp. FS406-22					5.502942e-03
|-- Methanomicrobia					
| `-- Methanosarcinales					
| `-- Methanosarcinaceae					
| `-- Methanosarcina					-9.402187e-02
`-- Thermococci					
`-- Thermococcales					
`-- Thermococcaceae					
`-- Pyrococcus					
`-- Pyrococcus horikoshii					-1.392597e-02
`-- Thaumarchaeota					
`--					
`-- Cenarchaeales					-7.421010e-02
`-- Cenarchaeaceae					-9.927388e-05
`-- Cenarchaeum					-4.973676e-04
`-- Cenarchaeum symbiosum					-1.499730e-04
-- Bacteria			-1.940299e-03		
|-- Acidobacteria					-3.363769e-02
|-- Aquificae		3.075914e-03	3.509304e-02		5.210773e-02
| `-- Aquificae		7.253295e-07	6.192598e-06		7.093965e-06
| `-- Aquificales			9.111036e-06		6.084547e-05
| `-- Desulfurobacteriaceae					1.386234e-02
| `-- Thermovibrio					2.136388e-04
| `-- Thermovibrio ammonificans					3.112958e-05
|-- Actinobacteria					
`-- Actinobacteria					
|-- Actinomycetales	-6.754544e-02		-2.907168e-02		-3.660475e-02
| |-- Actinomycetaceae					
| | `-- Arcanobacterium					1.954860e-02
| | `-- Arcanobacterium haemolyticum					9.238259e-05
| |-- Corynebacteriaceae					
| | `-- Corynebacterium					
| | |-- Corynebacterium aurimucosum					
| | |-- Corynebacterium efficiens					-1.901142e-01
| | `-- Corynebacterium glutamicum					-1.120029e-02
| |-- Frankiaceae					-2.278383e-02
| | `-- Frankia					-7.450726e-05
| | `-- Frankia alni					-2.013104e-01
| |-- Microbacteriaceae					-5.940775e-03
| |-- Micrococcaceae					
| | |-- Arthrobacter					
| | | |-- Arthrobacter arilaitensis					-3.732643e-03
| | | `-- Arthrobacter chlorophenolicus					4.338583e-02
| | `-- Rothia					
| | `-- Rothia mucilaginosa					-5.007777e-02
| |-- Mycobacteriaceae			-7.757990e-03		-5.261959e-02
| | `-- Mycobacterium			-3.554042e-04		-1.057588e-03
| | |-- Mycobacterium ulcerans		-3.532315e-02	-1.512812e-02		-9.259458e-02
| | `-- Mycobacterium sp. Spyr1					1.312925e-02
| |-- Nocardiaceae					-1.518301e-02
| | `-- Rhodococcus			-3.372657e-02		
| | `-- Rhodococcus jostii					-1.730418e-03
| |-- Nocardiopsaceae					
| | `-- Nocardiopsis					4.386254e-02
| | `-- Nocardiopsis dassonvillei					4.911622e-05
| |-- Promicromonosporaceae					3.419970e-02
| | `-- Xylanimonas					1.536712e-05
| |-- Pseudonocardiaceae					
| | `-- Saccharopolyspora					-1.297508e-02
| | `-- Streptomycetaceae					-4.758236e-02
| | `-- Streptomyces					-7.140394e-04
| | `-- Streptomyces scabiei					-3.286281e-02
| |-- Bifidobacteriales					
| | `-- Bifidobacteriaceae					
| | `-- Bifidobacterium					
| | `-- Bifidobacterium adolescentis					-1.337848e-02
| `-- Coriobacteriales					2.776751e-02
| `-- Coriobacteriaceae					1.819831e-04



| | | |-- Rhodobacterales								
| | | | `-- Rhodobacteraceae								
| | | | |-- Roseobacter							-6.929183e-02	-1.586300e-01
| | | | | `-- Roseobacter denitrificans							-3.135959e-04	-1.631854e-03
| | | | `-- Dinoroseobacter							-2.308099e-02	8.153308e-03
| | | | | `-- Dinoroseobacter shibae							-2.396911e-04	-1.077904e-03
| | | |-- Rhodospirillales							-2.396911e-04	-5.900223e-02
| | | | `-- Acetobacteraceae							-4.086096e-02	-3.489273e-02
| | | | |-- Acetobacter							-5.453880e-03	-1.077904e-03
| | | | |-- Gluconobacter							-6.308099e-02	-6.898738e-02
| | | | | `-- Gluconobacter oxydans							-2.396911e-04	-1.077904e-03
| | | | `-- Gluconacetobacter							-4.550520e-02	-4.461402e-04
| | | | | `-- Gluconacetobacter diazotrophicus							-4.461402e-04	
| | | `-- Sphingomonadales								
| | | | `-- Sphingomonadaceae								
| | | | | `-- Sphingopyxis							-2.037318e-02	
| | -- Betaproteobacteria								
| | | |-- Burkholderiales								
| | | | |-- Burkholderiaceae								
| | | | | |-- Burkholderia								
| | | | | | |-- Burkholderia ambifaria								2.480044e-02
| | | | | | |-- Burkholderia mallei							-1.693472e-02	-1.165139e-01
| | | | | | |-- Burkholderia pseudomallei							-5.146336e-02	-1.392108e-01
| | | | | | |-- Burkholderia rhizoxinica							-1.275355e-01	-2.696600e-01
| | | | | | |-- Burkholderia sp. CCGE1002								5.541674e-02
| | | | | | |-- Burkholderia sp. CCGE1003								7.162019e-03
| | | | | | `-- Burkholderia thailandensis								-2.298222e-02
| | | | | `-- Ralstonia								
| | | | | | `-- Ralstonia pickettii								2.848473e-02
| | | | |-- Comamonadaceae								
| | | | | |-- Alicycliphilus								9.414919e-03
| | | | | | `-- Alicycliphilus denitrificans								6.849592e-05
| | | | | `-- Polaromonas								-1.266691e-03
| | | | | | `-- Polaromonas naphthalenivorans								
| | | | `-- Oxalobacteraceae								
| | | | | `-- Candidatus Zinderia							-1.355157e-01	-2.320447e-01
| | | | | | `-- Candidatus Zinderia insecticola							-5.038534e-04	-7.124554e-04
| | | |-- Gallionellales								6.440292e-03
| | | | `-- Gallionellaceae								2.082755e-05
| | | |-- Neisseriales								-1.443947e-02
| | | | `-- Neisseriaceae								
| | | | | `-- Neisseria								
| | | | | | `-- Neisseria gonorrhoeae							-1.085558e-01	-1.620558e-01
| | | `-- Nitrosomonadales							1.067534e-02	
| | | | `-- Nitrosomonadaceae							1.490390e-05	
| | | | | `-- Nitrosospira								1.966773e-03
| | | | | | `-- Nitrosospira multiformis								5.334429e-05
| | -- Gammaproteobacteria								-4.904705e-03
| | | |-- Aeromonadales								
| | | | `-- Aeromonadaceae								
| | | | | `-- Aeromonas								-2.377393e-02
| | | |-- Alteromonadales								
| | | | `-- Shewanellaceae								
| | | | | `-- Shewanella								
| | | | | | |-- Shewanella baltica							-4.361515e-04	4.389922e-02
| | | | | | |-- Shewanella violacea								-2.930866e-03
| | | | | | `-- Shewanella woodyi							-2.539635e-01	
| | | |-- Chromatiales								2.173266e-02
| | | | `-- Ectothiorhodospiraceae								1.204713e-02
| | | | | `-- Thioalkalivibrio								9.919572e-03
| | | |-- Enterobacteriales							-2.219607e-02	-5.215810e-02
| | | | `-- Enterobacteriaceae							-6.953446e-04	-2.898716e-04
| | | | |-- Buchnera							8.942638e-02	1.248393e-01
| | | | | `-- Buchnera aphidicola							3.300004e-06	1.335786e-03
| | | | |-- Cronobacter								-8.895482e-02
| | | | | |-- Cronobacter sakazakii							1.255742e-02	
| | | | | `-- Cronobacter turicensis							-7.638727e-02	-5.962148e-02
| | | | |-- Erwinia								
| | | | | |-- Erwinia amylovora								-8.008089e-02
| | | | | `-- Erwinia pyrifoliae							-3.372900e-02	-3.868182e-02
| | | | |-- Escherichia							-4.105270e-02	-3.319002e-02
| | | | |-- Pantoea								
| | | | | |-- Pantoea ananatis								-2.495060e-02
| | | | | `-- Pantoea sp. At-9b								1.397057e-01
| | | | |-- Sodalis								-4.668960e-02
| | | | | `-- Sodalis glossinidius								-1.918559e-04
| | | | |-- Shigella							-1.666021e-02	-2.815831e-02
| | | | | |-- Shigella boydii							-2.874705e-02	
| | | | | `-- Shigella dysenteriae							-9.965514e-03	
| | | | |-- Wigglesworthia							7.107556e-03	3.481692e-03
| | | | | `-- Wigglesworthia glossinidia								1.437624e-05
| | | | |-- Candidatus Blochmannia								4.041529e-02
| | | | |-- Candidatus Riesia								-9.814730e-02
| | | | | `-- Candidatus Riesia pediculicola								-4.384943e-04
| | | | `-- Citrobacter								-5.662717e-04
| | | | | `-- Citrobacter koseri							-9.240591e-02	-1.215080e-01
| | | |-- Pasteurellales								
| | | | `-- Pasteurellaceae								
| | | | | `-- Haemophilus								
| | | | | | `-- Haemophilus influenzae								2.971605e-03
| | | |-- Pseudomonadales								-2.671661e-02
| | | | `-- Moraxellaceae								
| | | | |-- Acinetobacter								
| | | | | `-- Acinetobacter baumannii								-3.387477e-02
| | | | `-- Psychrobacter								7.865580e-03
| | | |-- Thiotrichales								
| | | | `-- Francisellaceae								
| | | | | `-- Francisella								
| | | | | | `-- Francisella tularensis								
| | | |-- Vibrionales								-1.572374e-02
| | | | `-- Vibrionaceae								-7.743855e-05
| | | | |-- Aliivibrio								
| | | | | `-- Aliivibrio fischeri								1.254021e-02
| | | | |-- Photobacterium								-3.958129e-02
| | | | | `-- Photobacterium profundum								-2.663203e-05
| | | | `-- Vibrio								-5.385216e-02
| | | | | |-- Vibrio cholerae								-3.855559e-02
| | | | | `-- Vibrio vulnificus								-1.007842e-02
| | | `-- Xanthomonadales							-6.643763e-02	-5.394485e-02
| | | | `-- Xanthomonadaceae							-1.283047e-03	-3.105384e-04
| | | | |-- Pseudoxanthomonas								3.358916e-02
| | | | | `-- Pseudoxanthomonas suwonensis								5.181536e-05
| | | | |-- Xanthomonas								-3.949454e-02
| | | | | |-- Xanthomonas axonopodis								-1.019246e-01

```
|   |   |           |   `-- Xanthomonas oryzae                                |-------------|-------------|-------------|-------------|-7.243194e-02|
|   |   |           `-- Xylella                                                |-------------|-------------|-7.166060e-02|-------------|-2.906261e-02|
|   |   |               `-- Xylella fastidiosa                                 |-------------|-------------|-2.864412e-04|-------------|-1.356959e-04|
|   |   |-- Epsilonproteobacteria                                              |-------------| 8.407714e-03|-------------|-------------| 2.143582e-02|
|   |   |   |-- Campylobacterales                                              |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |-- Campylobacteraceae                                         |-------------| 1.201179e-02| 1.142989e-02|-------------| 4.126519e-03|
|   |   |   |   |   |-- Arcobacter                                             |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |   |   `-- Arcobacter butzleri                                |-------------|-------------|-------------|-4.895792e-02|-------------|
|   |   |   |   |   `-- Campylobacter                                          |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |       |-- Campylobacter hominis                              |-------------|-------------|-------------|-------------|-7.873609e-02|
|   |   |   |   |       `-- Campylobacter jejuni                               |-------------|-------------|-------------|-------------| 8.783505e-04|
|   |   |   |   `-- Helicobacteraceae                                          |-------------|-------------|-------------|-------------|-------------|
|   |   |   |       |-- Helicobacter                                           |-------------|-------------|-------------|-------------|-------------|
|   |   |   |       |   `-- Helicobacter acinonychis                           |-------------|-------------|-------------|-------------|-3.423503e-02|
|   |   |   |       `-- Sulfuricurvum                                          |-------------|-------------|-------------|-------------| 4.336985e-02|
|   |   |   |           `-- Sulfuricurvum kujiense                             |-------------|-------------|-------------|-------------| 2.883400e-05|
|   |   |   `-- Legionellales                                                  |-------------|-------------|-------------|-------------|-------------|
|   |   |       `-- Coxiellaceae                                               |-------------|-------------|-1.905365e-02|-------------|-1.411842e-01|
|   |   |           `-- Coxiella                                               |-------------|-------------|-7.133419e-05|-------------|-4.987590e-04|
|   |   |               `-- Coxiella burnetii                                  |-------------|-------------|-2.859376e-05|-------------|-1.952051e-03|
|   |   `-- Deltaproteobacteria                                                |-------------|-------------|-------------|-------------|-------------|
|   |       |-- Desulfobacterales                                              |-------------|-------------|-------------|-------------|-------------|
|   |       |   `-- Desulfobacteraceae                                         |-------------|-------------|-------------|-------------|-------------|
|   |       |       `-- Desulfatibacillum                                      |-------------|-------------|-2.640722e-01|-------------|-------------|
|   |       |           `-- Desulfatibacillum alkenivorans                     |-------------|-------------|-1.118605e-03|-------------|-------------|
|   |       |-- Desulfuromonadales                                             |-------------|-------------|-------------|-------------|-------------|
|   |       |   |-- Geobacteraceae                                             |-------------|-------------|-------------|-------------| 1.871814e-03|
|   |       |   |   `-- Geobacter                                              |-------------|-------------|-------------|-------------| 1.720142e-04|
|   |       |   |       `-- Geobacter lovleyi                                  |-------------|-------------|-------------|-------------| 1.448883e-02|
|   |       |   `-- Pelobacteraceae                                            |-------------|-------------|-------------|-------------|-------------|
|   |       |       `-- Pelobacter                                             |-------------|-------------|-------------|-------------|-------------|
|   |       |           `-- Pelobacter propionicus                             |-------------|-------------|-------------|-------------|-4.534561e-02|
|   |       |-- Myxococcales                                                   |-------------|-------------|-------------|-------------|-------------|
|   |       |   |-- Myxococcaceae                                              |-------------|-------------|-------------|-------------|-------------|
|   |       |   |   `-- Myxococcus                                             |-------------|-------------|-------------|-------------|-1.794829e-01|
|   |       |   |       `-- Myxococcus xanthus                                 |-------------|-------------|-------------|-------------|-5.412304e-04|
|   |       |   `-- Polyangiaceae                                              |-------------|-------------|-------------|-------------|-1.317521e-01|
|   |       |       `-- Sorangium                                              |-------------|-------------|-------------|-------------|-6.381520e-04|
|   |       |           `-- Sorangium cellulosum                               |-------------|-------------|-------------|-------------|-2.693810e-04|
|   |       `-- Syntrophobacterales                                            |-------------|-------------|-------------|-------------|-3.880854e-03|
|   |           `-- Syntrophaceae                                              |-------------|-------------|-------------|-------------|-6.533699e-02|
|   |               `-- Syntrophus                                             |-------------|-------------|-------------|-------------|-6.352554e-04|
|   |                   `-- Syntrophus aciditrophicus                          |-------------|-------------|-------------|-------------|-2.206680e-05|
|   |-- Firmicutes                                                             |-------------|-------------|-------------|-------------| 3.867544e-04|
|   |   |-- Bacilli                                                            |-------------|-------------|-------------|-------------|-------------|
|   |   |   |-- Bacillales                                                     |-------------|-4.348195e-03|-4.252978e-02|-------------|-------------|
|   |   |   |   |-- Alicyclobacillaceae                                        |-------------|-------------|-3.498131e-02|-------------|-------------|
|   |   |   |   |-- Bacillaceae                                                |-------------|-8.536246e-03|-2.129847e-02|-------------|-1.194968e-02|
|   |   |   |   |   |-- Bacillus                                               |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |   |   |-- Bacillus cereus                                    |-------------|-------------|-------------|-------------|-2.970616e-02|
|   |   |   |   |   |   |-- Bacillus megaterium                                |-------------|-------------|-3.825868e-02|-7.692505e-02|-7.692648e-02|
|   |   |   |   |   |   `-- Bacillus thuringiensis                             |-------------|-------------|-------------|-------------|-1.216467e-02|
|   |   |   |   |   |-- Anoxybacillus                                          |-------------|-------------|-------------|-------------|-1.063135e-01|
|   |   |   |   |   |   `-- Anoxybacillus flavithermus                         |-------------|-------------|-------------|-------------|-5.661120e-04|
|   |   |   |   |   |-- Geobacillus                                            |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |   |   `-- Geobacillus thermodenitrificans                    |-------------|-------------|-------------|-------------|-3.061633e-02|
|   |   |   |   |   `-- Lysinibacillus                                         |-------------|-------------|-------------|-------------|-3.353529e-02|
|   |   |   |   |       `-- Lysinibacillus sphaericus                          |-------------|-------------|-------------|-------------|-7.175666e-05|
|   |   |   |   |-- Listeriaceae                                               |-------------|-------------|-------------|-------------| 1.884347e-02|
|   |   |   |   |   `-- Listeria                                               |-------------|-------------|-------------|-------------| 4.355847e-07|
|   |   |   |   |-- Staphylococcaceae                                          |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |   `-- Macrococcus                                            |-4.943999e-02|-------------|-------------|-------------|-------------|
|   |   |   |   |       `-- Macrococcus caseolyticus                           |-2.920222e-05|-------------|-------------|-------------|-------------|
|   |   |   |   `--                                                            |-------------|-------------|-------------|-------------|-------------|
|   |   |   |       `-- Exiguobacterium                                        |-------------|-------------|-------------|-------------| 8.111660e-02|
|   |   |   |           `-- Exiguobacterium sibiricum                          |-------------|-------------|-------------|-------------| 4.407049e-03|
|   |   |   `-- Lactobacillales                                                | 2.189671e-02|-------------|-------------|-------------| 2.508553e-03|
|   |   |       |-- Lactobacillaceae                                           |-------------|-------------|-------------|-------------| 7.199953e-03|
|   |   |       |   `-- Lactobacillus                                          |-------------|-------------|-------------|-------------|-------------|
|   |   |       |       |-- Lactobacillus johnsonii                            |-------------|-------------|-------------|-------------| 1.317637e-02|
|   |   |       |       `-- Lactobacillus plantarum                            |-------------|-------------|-2.612974e-02|-------------|-7.778998e-02|
|   |   |       |-- Leuconostocaceae                                           |-------------|-------------|-------------|-------------|-------------|
|   |   |       |   `-- Leuconostoc                                            |-------------|-------------|-------------|-------------|-------------|
|   |   |       |       `-- Leuconostoc mesenteroides                          |-------------|-------------|-------------|-------------| 2.142337e-02|
|   |   |       `-- Streptococcaceae                                           |-------------|-------------| 1.765274e-03|-------------|-------------|
|   |   |           |-- Lactococcus                                            |-------------|-------------|-------------|-------------| 3.155728e-02|
|   |   |           |   `-- Lactococcus lactis                                 |-------------|-------------|-------------|-------------| 2.611579e-05|
|   |   |           `-- Streptococcus                                          | 2.208334e-02|-------------|-------------|-------------|-------------|
|   |   |-- Clostridia                                                         | 3.643546e-02|-------------|-------------|-------------| 2.304012e-02|
|   |   |   |-- Clostridiales                                                  |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |-- Clostridiaceae                                             |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   |   `-- Alkaliphilus                                           |-------------|-------------|-------------|-2.471905e-01|-------------|
|   |   |   |   |-- Eubacteriaceae                                             |-------------|-------------|-------------|-------------|-2.860454e-03|
|   |   |   |   |   `-- Eubacterium                                            |-------------|-------------|-------------|-------------|-1.686161e-05|
|   |   |   |   |       `-- Eubacterium eligens                                |-------------|-------------| 6.211680e-02|-------------|-------------|
|   |   |   |   |-- Clostridiales Family XI. Incertae Sedis                    |-------------|-------------|-------------|-------------| 5.487347e-03|
|   |   |   |   |-- Clostridiales Family XVIII. Incertae Sedis                 |-------------|-------------|-------------|-------------|-4.889729e-02|
|   |   |   |   |   `-- Symbiobacterium                                        |-------------|-------------|-------------|-------------|-1.504438e-04|
|   |   |   |   |       `-- Symbiobacterium thermophilum                       |-------------|-------------|-------------|-------------|-1.983867e-04|
|   |   |   |   |-- Peptococcaceae                                             |-------------|-------------|-------------|-2.875267e-03|-------------|
|   |   |   |   |   |-- Desulfitobacterium                                     |-------------|-------------|-------------|-------------|-8.987518e-03|
|   |   |   |   |   `-- Thermincola                                            |-------------|-------------|-------------|-1.188084e-02|-------------|
|   |   |   |   |       `-- Thermincola potens                                 |-------------|-------------|-------------|-4.112690e-05|-------------|
|   |   |   |   `-- Syntrophomonadaceae                                        |-------------|-------------|-------------|-------------|-------------|
|   |   |   |       `-- Syntrophomonas                                         |-------------|-------------|-------------|-1.634025e-02|-------------|
|   |   |   |           `-- Syntrophomonas wolfei                              |-------------|-------------|-------------|-4.619094e-05|-------------|
|   |   |   |-- Halanaerobiales                                                |-------------|-------------|-------------|-------------|-------------|
|   |   |   |   `-- Halobacteroidaceae                                         |-------------|-------------|-------------|-3.114127e-02|-------------|
|   |   |   |       `-- Acetohalobium                                          |-------------|-------------|-------------|-8.612269e-06|-------------|
|   |   |   |           `-- Acetohalobium arabaticum                           |-------------|-------------|-------------|-1.214624e-05|-------------|
|   |   |   `-- Thermoanaerobacterales                                         |-------------|-------------|-------------|-------------| 7.189466e-03|
|   |   |       |-- Thermoanaerobacteraceae                                    |-------------|-------------|-------------|-------------|-------------|
|   |   |       |   `-- Thermoanaerobacter                                     |-------------|-------------|-------------|-------------| 4.433427e-03|
|   |   |       |       `-- Thermoanaerobacter mathranii                       |-------------|-------------|-------------|-3.526229e-02|-------------|
|   |   |       `-- Thermoanaerobacterales Family III. Incertae Sedis          |-------------|-------------|-------------|-------------| 4.587067e-02|
|   |   `-- Negativicutes                                                      |-------------|-------------|-------------|-------------| 1.999591e-02|
|   |       `-- Selenomonadales                                                |-------------|-------------|-------------|-------------| 2.579515e-05|
|   |-- Chlorobi                                                               |-------------|-------------|-------------|-------------|-------------|
|   |   `-- Chlorobia                                                          |-------------|-------------|-------------|-------------|-------------|
|   |       `-- Chlorobiales                                                   |-------------|-------------|-------------|-------------|-------------|
|   |           `-- Chlorobiaceae                                              |-------------|-------------|-------------|-------------|-------------|
|   |               |-- Chlorobaculum                                          |-------------|-------------|-------------|-------------|-------------|
```

			`-- Chlorobaculum tepidum			-9.123173e-03
		`-- Chlorobium				
		`-- Chlorobium chlorochromatii		-4.915972e-01	-1.696634e-02	
	`-- Chloroflexi					
		`-- Chloroflexi		-2.979740e-03		
			`-- Chloroflexales			
			`-- Chloroflexaceae			
			`-- Chloroflexus	-3.609203e-01		
	`-- Dehalococcoidetes					1.758129e-02
	`-- Thermomicrobia					
		`-- Sphaerobacterales				3.402833e-02
			`-- Sphaerobacteraceae			6.606473e-05
				`-- Sphaerobacter		5.090737e-06
					`-- Sphaerobacter thermophilus	3.294160e-05
`-- Chlamydiae						
	`-- Chlamydiae					
		`-- Chlamydiales				
			`-- Chlamydiaceae			
				|-- Chlamydia		
					`-- Chlamydia trachomatis	1.675127e-02
				`-- Chlamydophila		
					|-- Chlamydophila caviae	1.832506e-02
					`-- Chlamydophila felis	-9.888075e-02
`-- Deferribacteres						5.191689e-02
	`-- Deferribacteres					1.195181e-05
		`-- Deferribacterales				1.810409e-04
			`-- Deferribacteraceae			3.273859e-06
`-- Spirochaetes						-6.638125e-03
	`-- Spirochaetes					
		`-- Spirochaetales				
			`-- Spirochaetaceae		-1.232745e-02	-2.239349e-02
				|-- Borrelia	-5.674992e-02	-4.258764e-03
						-- Borrelia afzelii	-8.674963e-02	
						-- Borrelia bavariensis		-8.936955e-02
					`-- Borrelia burgdorferi	-1.545670e-03	-1.507201e-02
				`-- Treponema			
					`-- Treponema pallidum	-2.504165e-02	
`-- Synergistetes						1.353586e-02
	`-- Synergistia					1.023724e-04
		`-- Synergistales				2.685773e-05
			`-- Synergistaceae			2.167862e-04
`-- Tenericutes				-4.095060e-02		
	`-- Mollicutes					
			-- Mycoplasmatales	-2.503447e-01		
			`-- Mycoplasmataceae	-2.330382e-04		
				|-- Mycoplasma		
						-- Mycoplasma hyopneumoniae	-3.263517e-02	-1.084858e-01
					`-- Mycoplasma synoviae		-1.071024e-01
			`-- Ureaplasma		-2.747768e-01	-4.113904e-01	-1.428218e-01	-4.913782e-01
		`-- Entomoplasmatales			-1.300034e-01	-3.368641e-01		-4.559588e-01
			`-- Entomoplasmataceae		-7.810073e-04	-5.514984e-04		-6.859503e-04
			`-- Mesoplasma		-1.329830e-04	-3.073947e-05		-5.463459e-04
				`-- Mesoplasma florum	-6.635481e-09			-4.405472e-04
		`-- Acholeplasmatales				8.105822e-03		5.759166e-03
			`-- Acholeplasmataceae			9.200349e-06		2.484447e-05
				`-- Candidatus Phytoplasma				
					`-- Aster yellows witches'-broom phytoplasma	1.638140e-03	3.942714e-02		1.253411e-02
`-- Thermotogae								4.544117e-03
	`-- Thermotogae							1.451070e-05
		`-- Thermotogales						1.876103e-05
			`-- Thermotogaceae					4.821732e-06
				`-- Thermotoga				
					`-- Thermotoga neapolitana			-3.548564e-02
`-- Verrucomicrobia								
	`--							
		`-- Methylacidiphilales						-1.143331e-01
			`-- Methylacidiphilaceae					
				`-- Methylacidiphilum				-6.844453e-04
					`-- Methylacidiphilum infernorum			-4.898530e-04
	`--							
		`--						
			`--					
				`-- Thermobaculum				4.011630e-04
					`-- Thermobaculum terrenum			2.642266e-05
`-- Viruses								
	`--							
		`--						
			`--					
				`-- Microviridae			-1.993299e-01		-1.777363e-01
					`-- Chlamydiamicrovirus		-8.928895e-04		-3.686671e-04
						`-- Chlamydia phage CPAR39	-3.626180e-04		-1.946828e-03

Table 4.5: Lasso coefficients for precision (order) (5'+3')

The table shows the coefficients produced by performing the lasso on the
precision (order) model for exact matching gene start and stop.
The coefficients for each program are for the model chosen by the
'one-standard-error' rule (see section 3.4.2).
See also figure 4.7 for more information about values of the
shrinkage factors.

Coefficients	MED	Glimmmer	GeneMark.hmm	GeneMarkS	Prodigal
Intercept	6.198417e-01	7.859934e-01	7.103990e-01	7.200063e-01	9.617522e-01
GC	-3.940345e-01	-1.756062e-01	-7.252417e-02	-6.810346e-02	-4.062757e-02
Length (bp)	5.041270e-08	1.708885e-08	1.215465e-08	1.055431e-08	3.456334e-08
I(translation table = 11)		-8.279228e-02	-7.761006e-03		-2.475252e-01
Taxonomy					
|-- Archaea			1.381796e-02		
| |-- Crenarchaeota		2.137984e-03			
| | \`-- Thermoprotei		2.586347e-05			
| | |-- Acidilobales		1.258099e-01			
| | |-- Desulfurococcales					1.378190e-02
| | \`-- Thermoproteales		1.205234e-02			8.930091e-03
| |-- Euryarchaeota					
| | |-- Archaeoglobi		2.137163e-03			2.214998e-02
| | | \`-- Archaeoglobales		6.293247e-06			
| | |-- Halobacteria			4.310870e-02		-6.791253e-03
| | | \`-- Halobacteriales			6.204500e-06		-9.642075e-06
| | |-- Methanococci					3.601684e-02
| | | \`-- Methanococcales					1.715405e-04
| | \`-- Methanomicrobia					
| | |-- Methanomicrobiales					2.419117e-02
| | \`-- Methanosarcinales		-7.411629e-02			-4.086766e-02
| \`-- Thaumarchaeota					
| | \`--					
| | \`-- Cenarchaeales					-4.501735e-02
|-- Bacteria		-2.342462e-02	-2.323770e-04		
| |-- Acidobacteria	-1.951458e-02	-1.269185e-02			-6.509258e-02
| | \`-- Solibacteres					-7.028382e-02
| | \`-- Solibacterales					-2.015995e-03
| |-- Aquificae	4.509354e-02	1.025302e-01	4.380457e-02		8.845450e-02
| | \`-- Aquificae	6.540106e-05	4.137551e-05			7.603938e-05
| | \`-- Aquificales	3.450844e-05	1.087041e-04	1.860856e-04		1.796108e-04
| |-- Actinobacteria	-3.201184e-02		-5.717854e-03		
| | \`-- Actinobacteria			-1.708122e-06		
| | |-- Actinomycetales	-2.950209e-03	-2.544241e-02	-3.937084e-02		-7.630494e-02
| | |-- Acidimicrobiales					2.373360e-02
| | |-- Bifidobacteriales					
| | \`-- Coriobacteriales					8.791054e-02
| |-- Bacteroidetes			9.760529e-03		
| | |-- Bacteroidia	5.247609e-04				-2.680333e-02
| | | \`-- Bacteroidales					-1.351364e-03
| | |-- Cytophagia		2.159305e-02	4.488301e-02		
| | | \`-- Cytophagales		6.352032e-05	6.460628e-05		
| | |-- Flavobacteria	1.419461e-02	1.030760e-01	5.508099e-02		3.707401e-02
| | | \`-- Flavobacteriales		2.076020e-04	8.855709e-04		3.396391e-04
| | \`-- Sphingobacteria	-5.652908e-03				-1.189205e-01
| | \`-- Sphingobacteriales	-1.657476e-05				-7.173060e-05
| |-- Cyanobacteria		1.838583e-02			
| | |-- Gloeobacteria					-9.168813e-03
| | | \`-- Gloeobacterales					-1.008655e-05
| | \`--					
| | |-- Nostocales		-1.775315e-02			-6.134117e-02
| | |-- Oscillatoriales		-2.474405e-01			-1.050580e-01
| | \`-- Prochlorales		5.092107e-02	4.549607e-02		2.953682e-02
| |-- Deinococcus-Thermus					
| | \`-- Deinococci					
| | |-- Deinococcales					-2.760285e-02
| | \`-- Thermales	3.764535e-02	2.154964e-02	2.986007e-02		6.353794e-02
| |-- Dictyoglomi		2.465455e-02			8.114052e-02
| | \`-- Dictyoglomia		5.255672e-06			1.607296e-05
| | \`-- Dictyoglomales		1.893252e-04			6.497589e-05
| |-- Elusimicrobia		-1.751549e-01			-4.386661e-02
| | \`-- Elusimicrobia		1.953088e-01			1.270717e-01
| | \`-- Elusimicrobiales		4.313192e-05			1.149656e-04
| |-- Fusobacteria					1.091862e-01
| | \`-- Fusobacteria					4.925449e-05
| |-- Planctomycetes	-1.149305e-01	-5.544651e-02	-1.712404e-02		-1.495610e-02
| | \`-- Planctomycetacia	-9.683649e-05	-8.073925e-07	-6.540762e-06		-3.184810e-05
| | \`-- Planctomycetales	-4.775645e-05	-4.115789e-05	-3.849757e-04		-1.287552e-05
| |-- Proteobacteria					
| | |-- Alphaproteobacteria					
| | | |-- Caulobacterales		1.142383e-02			2.252780e-02
| | | |-- Rickettsiales		-3.439144e-03			-2.789079e-02
| | | |-- Rhizobiales	-6.810810e-02	-3.859194e-02	-3.756417e-02		-3.907321e-02
| | | |-- Rhodobacterales		3.826131e-02			
| | | |-- Rhodospirillales	-8.989227e-03	-8.620218e-02	-6.844653e-02		-1.327574e-01
| | | \`-- Sphingomonadales		1.097186e-02			
| | |-- Betaproteobacteria					
| | | |-- Gallionellales					7.785874e-02
| | | |-- Neisseriales			-5.326218e-03		-2.421982e-02
| | | \`-- Methylophilales		2.723303e-02			4.678408e-02
| | |-- Gammaproteobacteria					-4.732330e-03
| | | |-- Alteromonadales		5.040836e-03			1.209202e-02
| | | |-- Cardiobacteriales					1.721611e-02
| | | |-- Chromatiales					3.698585e-02
| | | |-- Enterobacteriales		-9.545455e-03	-3.107829e-03		-3.096568e-02
| | | |-- Pasteurellales	1.271403e-01	6.720409e-03			1.027664e-02
| | | |-- Pseudomonadales	-1.877518e-02	-1.857427e-03	-6.200533e-03		-6.827224e-02
| | | |-- Vibrionales	1.194954e-02	4.442276e-04			-7.443725e-04
| | | \`-- Xanthomonadales	-5.130199e-02		-8.486456e-02		-7.366058e-02
| | |-- Epsilonproteobacteria	5.973260e-02	9.262830e-02	6.214921e-02		1.105883e-01
| | | \`-- Legionellales					-4.563310e-02
| | \`-- Deltaproteobacteria	-1.661603e-02	-1.159362e-02			
| | |-- Desulfarculales					3.730503e-02

| | |-- Desulfobacterales			-6.534719e-03		
| | |-- Myxococcales		-8.558819e-02			-6.480563e-02
| | `-- Syntrophobacterales		-1.063590e-02			-2.744794e-02
| |-- Firmicutes					2.098199e-02
| | |-- Bacilli					
| | | |-- Bacillales		-1.159546e-02	-2.929589e-02		
| | | `-- Lactobacillales	5.550564e-02	3.017443e-02			1.254904e-02
| | |-- Clostridia	4.640897e-02				3.643809e-02
| | | |-- Halanaerobiales					5.394303e-03
| | | |-- Natranaerobiales		3.124644e-02			5.285054e-03
| | | `-- Thermoanaerobacterales					1.039282e-02
| | `-- Negativicutes	7.669500e-04				7.480894e-02
| | `-- Selenomonadales					5.325839e-04
| |-- Chloroflexi					
| | |-- Chloroflexi			-2.287103e-02	-3.759211e-02	-3.701578e-03
| | | `-- Chloroflexales	-3.161151e-01				
| | |-- Dehalococcoidetes	9.233720e-02	4.809408e-02			9.347151e-02
| | `-- Thermomicrobia	4.082359e-02	1.072328e-02			3.541805e-02
| | |-- Sphaerobacterales					6.816529e-02
| | `-- Thermomicrobiales		5.711134e-02			
| |-- Chlamydiae	6.085276e-03	9.954634e-03			3.902473e-02
| | `-- Chlamydiae	1.063859e-05	2.450981e-05			2.966088e-04
| | `-- Chlamydiales		1.908981e-04			4.235058e-04
| |-- Deferribacteres		2.517610e-02	2.539449e-03		9.396333e-02
| | `-- Deferribacters		9.168957e-05	7.149759e-07		3.781247e-05
| | `-- Deferribacterales		1.067084e-04	1.155540e-05		7.379781e-04
| |-- Spirochaetes	-8.375004e-02	-6.906509e-03	-8.859564e-03		-3.253427e-02
| | `-- Spirochaetes	-8.165502e-06	-3.236625e-06	-2.029746e-05		-3.266107e-05
| | `-- Spirochaetales	-8.704659e-06	-8.648774e-05	-1.789341e-04		-7.139956e-05
| |-- Synergistetes					8.741674e-02
| | `-- Synergistia					2.001806e-06
| | `-- Synergistales					1.429628e-04
| |-- Tenericutes					
| | `-- Mollicutes					
| | |-- Mycoplasmatales	-4.060041e-01	-3.756960e-02			-2.160275e-01
| | |-- Entomoplasmatales	-7.083619e-02	-4.130502e-01	-3.799067e-01		-5.223328e-01
| | `-- Acholeplasmatales	9.279248e-02	6.351571e-02	3.640822e-02		7.520244e-02
| |-- Thermotogae	1.050977e-01				5.639537e-02
| | `-- Thermotogae	7.712420e-05				3.712116e-04
| | `-- Thermotogales					5.326619e-04
| |-- Verrucomicrobia					
| | |-- Opitutae					
| | | `-- Puniceicoccales			-2.366975e-01		7.251285e-03
| | `--					
| | `-- Methylacidiphilales					-1.658563e-02
| |-- Chrysiogenetes					3.323105e-02
| `-- Gemmatimonadetes					-2.325203e-02
| `-- Gemmatimonadetes					-5.227854e-05
| `-- Gemmatimonadales					-1.428497e-05
`-- Viruses					
`--					
`--					
`-- Caudovirales	1.084429e-01				5.382195e-03

Table 4.6: Lasso coefficients for recall (order) (5'+3')

The table shows the coefficients produced by performing the lasso on the
recall (order) model for exact matching gene start and stop.
The coefficients for each program are for the model chosen by the
'one-standard-error' rule (see section 3.4.2).
See also figure 4.8 for more information about values of the
shrinkage factors.

Coefficients	MED	Glimmmer	GeneMark.hmm	GeneMarkS	Prodigal
Intercept	5.895970e-01	7.846097e-01	8.081797e-01	8.213316e-01	8.605134e-01
GC	-8.784900e-02	-1.859910e-01	-1.809166e-01	-1.803304e-01	-1.022846e-01
Length (bp)	4.968584e-08	1.170798e-09	6.721303e-09	1.279101e-09	2.299606e-08
I(translation table = 11)					-4.914220e-02
Taxonomy					
|-- Archaea			1.170425e-02		
| |-- Crenarchaeota					
| | `-- Thermoprotei					
| | `-- Desulfurococcales			-1.984607e-02		
| |-- Euryarchaeota					
| | |-- Halobacteria			4.456351e-02		
| | |-- Methanococci					1.596498e-02
| | | `-- Methanococcales					1.367784e-05
| | `-- Methanomicrobia					
| | `-- Methanomicrobiales					7.076638e-03
| `-- Thaumarchaeota					
| `--					
| `-- Cenarchaeales					-7.667602e-02
`-- Bacteria					
|-- Acidobacteria					-2.894488e-02
|-- Aquificae		2.617077e-02	5.825298e-02		7.172467e-02
| `-- Aquificae		3.638087e-06	2.738705e-05		1.317464e-05
| `-- Aquificales		3.943023e-05	1.994634e-04		1.042354e-05
|-- Actinobacteria					
| `-- Actinobacteria					
| |-- Actinomycetales	-8.225679e-02		-3.553304e-02		-4.548574e-02
| |-- Acidimicrobiales					2.161920e-03
| `-- Coriobacteriales					4.851780e-02
|-- Bacteroidetes			4.073857e-02		
| |-- Bacteroidia					-3.045025e-03
| | `-- Bacteroidales					-4.116064e-07
| |-- Flavobacteria		2.613204e-02	1.012146e-02		6.214584e-03
| | `-- Flavobacteriales			2.533564e-05		1.962273e-05
| `-- Sphingobacteria			-2.098928e-02		-7.835290e-02
| `-- Sphingobacteriales			-3.828993e-05		-2.304746e-04
|-- Cyanobacteria		1.630498e-02			-5.566388e-03
| |-- Gloeobacteria					
| | `-- Gloeobacterales					
| `--					
| |-- Nostocales			1.893598e-02		
| `-- Prochlorales			1.179323e-03		
|-- Deinococcus-Thermus					
| `-- Deinococci					
| |-- Deinococcales					-8.640717e-03
| `-- Thermales			3.987513e-02		4.717066e-02
|-- Dictyoglomi					1.727308e-02
|-- Elusimicrobia			4.633566e-02		5.070612e-02
|-- Fusobacteria					7.965306e-02
| `-- Fusobacteria					
| `-- Fusobacteriales					1.894468e-05
|-- Planctomycetes	-1.674571e-01		-9.036811e-02		-4.652164e-02
| `-- Planctomycetacia			-2.610802e-07		-4.367656e-04
| `-- Planctomycetales			-1.306984e-05		-5.482760e-05
|-- Proteobacteria					
| |-- Alphaproteobacteria					
| | |-- Caulobacterales					1.752547e-02
| | |-- Rhizobiales		-1.137749e-02	-1.536873e-02		-1.878847e-02
| | `-- Rhodospirillales	-9.161399e-03	-5.182487e-02	-4.056034e-02		-8.907496e-02
| |-- Betaproteobacteria					
| | |-- Gallionellales					3.081799e-02
| | |-- Neisseriales			-3.570150e-02		-4.987143e-02
| | |-- Methylophilales					1.093976e-02
| | `-- Nitrosomonadales			3.835984e-02		9.245651e-04
| |-- Gammaproteobacteria					
| | |-- Alteromonadales					1.473102e-02
| | |-- Chromatiales			8.609895e-03		4.007528e-02
| | |-- Enterobacteriales			-3.395216e-02		-5.649346e-02
| | |-- Pasteurellales	8.228354e-04		6.404054e-03		8.537536e-03
| | |-- Pseudomonadales	-8.756670e-03				-3.198833e-02
| | |-- Thiotrichales			2.184419e-02		4.050898e-03
| | |-- Vibrionales					-5.370655e-02
| | `-- Xanthomonadales	-3.125366e-02		-9.780245e-02		-9.868505e-02
| |-- Epsilonproteobacteria	3.232117e-03	2.373961e-02	1.653151e-02		3.713723e-02
| | `-- Legionellales					-5.238237e-02
| `-- Deltaproteobacteria					
| |-- Desulfobacterales			-2.615296e-03		
| |-- Myxococcales	-6.959773e-02				-2.948335e-03
| `-- Syntrophobacterales					-3.453219e-02
|-- Firmicutes					
| |-- Bacilli					
| | |-- Bacillales		-1.436121e-02	-5.815039e-02		
| | `-- Lactobacillales	4.643967e-02				1.773399e-02
| |-- Clostridia	4.926048e-02				3.559384e-02
| | `-- Thermoanaerobacterales					3.509662e-02
| `-- Negativicutes					4.416626e-02
|-- Chloroflexi					6.600844e-03
| |-- Chloroflexi		-2.342022e-02			
| | `-- Chloroflexales	-2.560400e-01				
| |-- Dehalococcoidetes					3.032884e-02
| `-- Thermomicrobia					
| `-- Sphaerobacterales					5.307200e-02
|-- Deferribacteres	1.404509e-03		1.479997e-02		7.069537e-02
| `-- Deferribacteres	1.281287e-06		1.160867e-06		

`-- Deferribacterales	8.121842e-05	6.032930e-05		
-- Spirochaetes	-4.318966e-02	-4.748680e-03		-2.604278e-02
`-- Spirochaetes	-8.345420e-04	-1.800315e-04		-2.110931e-05
`-- Spirochaetales	-9.917791e-03	-3.686857e-05		-1.245504e-03
-- Synergistetes				3.816506e-02
`-- Synergistia				1.240215e-04
-- Tenericutes	-5.618304e-02			
`-- Mollicutes				
|-- Mycoplasmatales	-2.712057e-01			-5.830470e-02
|-- Entomoplasmatales		-2.043364e-01	-3.959508e-01	-4.618286e-01
`-- Acholeplasmatales		2.085569e-02	5.514503e-02	2.903232e-02
-- Thermotogae	3.089000e-02			1.823862e-02
`-- Thermotogae	7.658436e-05			3.839167e-05
`-- Thermotogales	1.115128e-06			
`-- Verrucomicrobia				
`--				
`-- Methylacidiphilales				-1.182596e-01

Chapter 5

Discussion

In this chapter the results from chapter 4 is discussed. The discussion is spread over three sections. The first section (§ 5.1) discusses the results from § 4.1 in which various visualisations were performed on the data. The second section (§ 5.2) discusses the results from the statistical analysis performed in § 4.2. The last section (§ 5.3) discusses some other points of interest.

5.1 Visualisation of the data

5.1.1 Number of coding genes v. sequence length

All of the gene prediction programs predicts, on average, more genes than what is annotated. This means that there always will be an amount of genes that are falsely predicted as coding genes.

GeneMarkS predicts 899 genes per megabase which is similar to the number of genes per megabase that is annotated. However, GeneMarkS also has some false positive genes. This means that GeneMarkS will not be able to detect all of the annotated genes since some of these will be falsely discarded as non-coding.

It is thus possible to say that GeneMarkS appears to be slightly too conservative in its gene predictions.

Prodigal, on the other hand, predicts 914 genes per megabase. This is 18 genes more per megabase than the annotation. In addition, it seems that Prodigal achieves higher precision and recall than GeneMarkS. The result is that, by being slightly more liberal than GeneMarkS, Prodigal seems to be able to find more of the annotated genes.

It is also noticeable that the annotation, GeneMarkS and Prodigal have about 900 genes per megabase. From prior knowledge (Konstantinidis and Tiedje 2004) it is expected that the number of genes per megabase is about 900. Thus it seems that that GeneMarkS and Prodigal produces an amount of genes that seems to be within a range that is suitable, but it is important to notice that not all of the predicted genes are correct.

The second group of programs — GeneMark.hmm and Glimmer — produces around 950 genes per megabase. This seems to be too liberal

for achieving optimal performance with regard to precision and recall. However, these programs seems to have a smaller degree of spread in their precision and recall results when compared to the first group.

GeneMarkS is an improvement of GeneMark.hmm. This is supported by the observed prediction performance which is higher for GeneMarkS than GeneMark.hmm. This improvement seems to be at the cost of a higher degree of spread in the prediction performance results. However, it is important to note that GeneMarkS is self-training, while GeneMark.hmm is based on previously constructed models. This means that the extra variance might be due to the self-training step.

The third group consists of only one program: MED. This program predicts almost 1,100 genes per megabase. This seems to be too liberal and thus seem to result in poor performance when it comes to accurately matching genes. MED also has a high degree of spread in its results.

However, MED still achieves fairly good gene detection when looking at recall. This means that the poor gene match performance might be due to the method used for separation of the decision boundary in MED's prediction algorithm (see § 2.3.5). However, a higher gene matching performance might be achieved by using a different method for separation of the decision boundary.

5.1.2 Violin plots for precision and recall

The violin plots for gene detection show that all of the gene prediction programs performs well with regard to gene detection. This indicates that accurately detecting the majority of genes in a sequence is currently not a problem for any of the programs.

However, looking at the violin plots for gene matching shows that the accurate matching of genes is harder. All of the programs perform worse for gene matching than gene detection. In addition, the gene matching performance displays a higher degree of spread.

However, the proportion of sequences with precision and recall above 90% for Prodigal shows that Prodigal provides a good degree of improvement in gene matching performance when compared to earlier programs.

5.1.3 Accuracy of start and stop codon prediction

Looking at the plots for the start codon accuracy for genes with correctly predicted stops indicates that the accurate finding of start codons still is a problem for all of the programs. However, most of the mismatched genes are mismatched by a whole number of codons. In addition, the probability for predicting a start far away from the correct start is small for all programs. This indicates that the programs are not making predictions far from the truth, and thus only need to improve the accuracy of start codon calling in order to get more satisfactory results.

Looking at the plots for the stop codon accuracy for genes with correctly predicted starts indicates that finding the correct stop when the start is

found is almost a certain event for all programs. This means that finding the stop if the start is found is not a problem for any of the programs.

5.2 Statistical analysis

The lasso models for precision and recall generally achieves a low degree of fit. This might be improved by developing methods based on some 'features' that the sequences have, since it is expected that similar sequences performs similarly. However, the development of such features fall outside the scope of this thesis and is thus left as a challenge for future research. See § 6.2.5.

5.2.1 Technical problems

While it was known by the author of this thesis at the time of running the gene prediction programs that the genus *Mycoplasma* uses a different genetic code (11) than the normal (4) for prokaryotes, it was not known that the orders *Mycoplasmatales* and *Entomoplasmatales*, and the species *Candidatus Hodgkinia cicadicola* and *Candidatus Zinderia insecticola*, also utilizes this genetic code.

The result of this is that these organisms were run using the wrong genetic code. When using the wrong genetic code, the programs will try to use the codon UGA for potential stop codon. This leads to a high number of predicted genes with wrong stop codons. The result is that these genes will be predicted to be shorter than what is acutally the case.

The large number of genes with wrongly predicted stop will decrease the recall and precision, and as a result the coefficients are large since the programs will perform worse than expected for these sequences, when the wrong translation table is used. This makes the coefficients for the affected sequences wrong, and thus any further discussion will ignore the flawed results.

See table 5.1 for a list of species in the reference data set that utilizes the codon UGA for tryptophan, in addition to GenBank accession, and if the results for a given species is flawed or not.

In addition, the author of this thesis forgot to run the program ELPH as part of the Glimmer pipeline. Since this is used for RBS improvements, this will make Glimmer seem to perform worse with regard to gene matching than what might actually be the case. Any further discussion of Glimmer's gene matching performance will thus be based on these unimproved predictions produced by Glimmer without running ELPH as part of the pipeline.

5.2.2 Taxonomy

Reductive evolution Looking at the tables for precision (see table 4.3) and recall (see table 4.4) there are two species — *Anabaena azollae* and *Mycobacterium leprae* — that have large negative coefficients for precision,

Table 5.1: Species in the reference data set that utilizes the codon UGA for tryptophan.

Species	GenBank accession	Flawed results
Candidatus Hodgkinia cicadicola	NC_012960	Yes
Candidatus Zinderia insecticola	NC_014497	Yes
Mesoplasma florum	NC_006055	Yes
Mycoplasma agalactiae	NC_009497, NC_013948	No
Mycoplasma arthritidis	NC_011025	No
Mycoplasma bovis	NC_014760	No
Mycoplasma capricolum	NC_007633	No
Mycoplasma conjunctivae	NC_012806	No
Mycoplasma crocodyli	NC_014014	No
Mycoplasma fermentans	NC_014552, NC_014921	No
Mycoplasma gallisepticum	NC_004829	No
Mycoplasma genitalium	NC_000908	No
Mycoplasma hominis	NC_013511	No
Mycoplasma hyopneumoniae	NC_006360, NC_007332, NC_007295	No
Mycoplasma hyorhinis	NC_014448	No
Mycoplasma leachii	NC_014751	No
Mycoplasma mobile	NC_006908	No
Mycoplasma mycoides	NC_005364	No
Mycoplasma penetrans	NC_004432	No
Mycoplasma pneumoniae	NC_000912	No
Mycoplasma pulmonis	NC_002771	No
Mycoplasma synoviae	NC_007294	No
Ureaplasma parvum	NC_010503, NC_002162	Yes
Ureaplasma urealyticum	NC_011374	Yes

but no coefficients for recall. This is because these species has a high number of *pseudogenes*. *A. azollae* has a pseudogene to coding gene ratio about 45.7%, while *M. leprae* has a ratio of about 69.5%. Since the pseudogenes might look like coding genes, although they are not active any more, the precision for the programs will decrease since the pseudogenes will be predicted as coding.

While these are among organisms with the highest pseudogene to coding gene ratio, further investigation of the tables reveals several organisms with a high ratio. Some exemples are *Mycobacterium ulcerans* (18.8%), *Orientia tsutsugamushi* (31.7%), *Rickettsia massiliae* (42.5%), *Sodalis glossindius* (40%) and *Trichodesimum erythraeum* (14%).

The common thread between these species is that they are *symbionts*. Since a symbiotic lifestyle provides a more stable habitat for the organism, the selection pressure to keep unused genes decreases. The reduced selection pressure results in inactivation of genes. These genes are know as *pseudogenes*.

Over time the nonfuctional DNA will become removed, due to *deletional bias*, and thus the genome size becomes reduced. (Moran 2003)

In other words, a high number of pseudogenes might indicate that the organism has recently become a symbiont. A lower number indicates that the regions containing pseudogenes has been deleted, and thus the organism is later in the genome reducing phase. This is expected to be

observed in the precision for the gene prediction programs, since a high number of pseudogenes increases the number of false positives, and thus reduces the precision. Similarly, when the number of pseudogenes reduces, the number of false positives becomes smaller and the precision becomes higher.

When looking at the results in table 4.1, it is noticable that organisms with a high pseudogene ratio performs worse with regard to the gene matching performance for precision.

For many of the organisms in table 4.1 it is also evident that the gene prediction programs perform worse with regard to recall. The poor recall performance might indicate that the performance of the gene prediction programs become poor as a result of including pseudogenes in the training model. The result of this is that the predictions will be dominated by the pseudogenes, and thus the programs will fail to find many of the real genes, since these does not look like what the programs expect from the training model.

Expansive evolution A suitable extension of the hypothesis from the discussion of the results for reductive evolution is to observe worse performance for outliers over the line for the annotations.

However, looking at these outliers, such as for the different strains of *Prochlorococcus marinus* and the *Bacillus cereus* group, no indication of that the performance of the programs suffers when the gene density is high were found. This is probably due to that the gene density is not high enough for the gene prediction programs to start filtering out genes due to their short length.

Other points MED has large negative coefficients in both the full and order model for the order *Chloroflexales* This is because MED failed to produce prediction results for these sequences. The reason is that the program caused segmentation faults when run on these sequences. The reason for this is unknown.

5.2.3 GC content

From the coefficients in tables 4.3 and 4.4, it is observable that the performance becomes poorer when the GC content of the sequence increases. This is probably due to that a high GC content leads to a low abundance of stop codons and a high abundance of start codons. (Hyatt et al. 2010)

A low abundance of stop codons will lead to a large number of long ORFs that are not coding genes. These ORFs might as a result get selected as predicted coding genes. The result of this is lower precision. (Hyatt et al. 2010)

A high abundance of start codons will make it harder to accurately select the correct start codon in a ORF, and thus might lead to a worse

gene matching performance due to wrongly predicted start codons. (Hyatt et al. 2010)

The observed coefficients are moderate. But since the coefficients signifies a change in GC content of 100%, the difference in performance between a GC rich and GC poor sequence is small. This is probably due to that all of the programs look for the RBS for each potential ORF and thus is able to filter out the long non-coding ORFs. The RBS finding methods implemented in each program also aids the selection of the correct start codon and thus improves the gene matching performance.

It is also noticeable that the coefficients for GC content for MED are twice as large as the second largest coefficient. This might indicate the method used in MED is highly sensitive for GC content. The result is that the precision for MED drops more than the other gene prediction programs when the GC content gets high.

It is also noticeable that the effect of difference in GC content is more important than taxonomy. It is also more important than length unless the difference in sequence length is large (several megabases).

Again, the coefficients for precision are larger than the coefficients for recall.

5.2.4 Sequence length

The coefficents for sequence length are small for all of the programs. This might indicate that the performance is not highly dependent on the sequence length. However, the performance are much variable for short sequences, e.g., plasmids. This might indicate that the performance is more dependent on if a sequence is very short or not, and thus this measure might be more useful if it were based on if a sequence were longer or shorter than some suitable threshold.

5.2.5 Full model v. order model

Most of the things mentioned above for the full model also applies for the order models with some minor differences.

The first noticable difference is that the species undergoing reduction evolution are not detectable when using the order model.

The next noticable difference is that the sequences run with the wrong genetic code has larger coefficients in the order model that the full model. However, the coefficients for the genus *Mycoplasma*, for the order model, is large and negative for MED, while being small or zero for the other programs. This indicates that the performance of MED for species using a different genetic code suffers greatly. This is because MED assumes all sequences uses the same genetic code (11). It is worth noticing that this performance drop was not detectable using the full model.

5.3 Other points of interest

5.3.1 Program optimisation

Glimmer has a high degree of tweakablility. This means that by applying *domain knowledge* it might be possible to improve the gene prediction performance. The other programs, on the other hand, has a lower degree of tweakability and thus their performance can not potentially be improved by domain knowledge.

However, notice that this thesis is not about the optimisation of the settings used to run the various gene prediction programs. The main focus is on evaluation the gene prediction preformance of the various programs. This is done under the assumption that the programs use good default settings.

It might be the case that the good performance observed for Prodigal is due to Prodigal's more advanced RBS finder and not by using dynamic programming instead of Markov models. This means that by combining the more advanced RBS finder of Prodigal with Glimmer might achieve high performance, low spread and allow the usage of domain knowledge.

5.3.2 Rare features

For both the start and stop codon plots there are some peaks that are present for all of the programs. These peaks probably appear due to rare features.

None of the programs where able to handle rare features such as introns or the usage of selenocysteine and pyrrolysine as amino acids. However, this might be a hard problem to solve without implementing RNA secondary structure prediction into the gene prediction programs. This means that these edge cases might better be handled outside of the gene prediction programs. See also § 6.1.1.

Chapter 6

Future work

IN THIS CHAPTER various aspect related to future work will be described. The first half of the chapter proposes various ways that the gene prediction programs can be improved. The second half proposes various ways the methods used in this thesis can be improved.

6.1 Improvement of gene prediction programs

6.1.1 Rare features

As mentioned in § 5.3.2, none of the gene prediction programs deal with the more rare features releated to coding genes of the prokaryotes.

While rare in prokaryotes, introns still appear from time to time in prokaryotic sequences. These introns are mainly *archaeal introns* and can be detected using RNA secondary structure prediction. (Watanabe et al. 2002)

Another rare feature is the amino acids *selenocysteine* and *pyrrolysine*. Selenocysteine is known as the 21st amino acid, and appear by reassining the stop codon UGA to code for the amino acid selenocysteine depending on context. The assigning requires what is know as a SECIS element, which is a hairpin loop in the RNA located downstream of the codon UGA for *Bacteria* or in the 3′-untranslated region (UTR) for *Archaea* and *Eukarya*. (Y. Zhang et al. 2005)

Pyrrolysine is know as the 22nd amino acid, and appear by assining the stop codon UAG to code for the amino acid pyrrolysine depending on context. However, pyrrolysine does not require a PYLIS element similar to the SECIS element found when selenocysteine is used. Pyrrolysine is mainly present in some archaea, e.g., *Methanosarcina acetivorans*, but the bacterium *Desulfitobacterium hafniense* also uses pyrrolysine. In addition, *D. hafniense* also uses selenocysteine and, as of 2005, is the only organism known to utilize both selenocysteine and pyrrolysine. (Y. Zhang et al. 2005)

Looking at these rare features, the common tread is that they can be detected by looking at the RNA secondary structure. As a result it might be hard to handle detection of these features inside the gene prediction programs, without implementing RNA secondary structure prediction.

This means that these rare features might best be handled outside of the gene prediction programs.

6.1.2 Partial ORFs

As the number of prokaryotic draft genomes increases the need to handle draft genomes becomes apparent.

An example of a problem that appears when working on draft genomes is genes that land on contig boundries, i.e., first bit of the gene lands on the end of one contig, and last bit lands on the start of another contig. This is what know as *gene fragmentation* and is studied closer in Klassen and Currie (2012).

By not handling partial ORFs, it becomes harder to perform, e.g., phylogenetic analyses using draft genomes. On the other hand, partial ORFs can introduce biases, but in many cases allowing partial ORFs might be useful.

6.2 Improvement of analysis

6.2.1 Draft genomes

Next-generation sequencing has allowed for cheap, fast sequencing of prokaryotic genomes. Since these technologies produce short reads, the economical trade-off gives a data set that has a huge breadth, but lacks the depth. Since it might not be possible to complete prokaryotic genomes with only short reads, it is likely that a large number of genomes will be of draft quality in the near future. (Klassen and Currie 2012)

Inspired by this, an further study of prokaryotic gene prediction programs should include draft genomes since the number of draft genomes is likely to grow faster than the number of completed genomes. As a result, any improvement of the gene programs might only be possible by also studying draft genomes.

6.2.2 Database schema

The database schema used in this thesis (see § 3.1.1) is too simple to allow for more advanced biological analysises that what has been performed here. As an example, the current schema does not handle genes with introns, so in the current schema a gene containing an intron is stored as a gene with the start from the first exon and the end from the last exon. However, since none of the programs handle introns (see §§ 5.3.2 and 6.1.1), the effect of this on the analysis were minimal.

Another example of things one might want to handle is features other that just coding genes. The current database schema only contains, for the coding genes, information about start, stop and strand, in addition to external references for annotated genes, if available.

6.2.3 Robust statistics

Most of the statistical analysis in this thesis uses classical statistical estimators such as the mean. However, such estimators are quite sensitive to outliers in the data, and given the noisy data used in this thesis these estimators might give suboptimal results. By using robust estimators, such as the median, better results might be achieved.

6.2.4 Bayesian approach

It is also possible to view the performance measures precision, recall and F_β-score is a probabilistic light. (Goutte and Gaussier 2005) This might give better results in the analysis by taking into account the uncertainty in the data, but this approach was not used in this master thesis. This is because of the lack of time to further develop this to a level usable for the analysis performed in this thesis.

6.2.5 Alternative sequence measures

Looking at the fit for the models for precision and recall that were shrunk with the lasso, it is noticable that the fit of the models falls in the lower end of the scale. This fit might be improved by developing, e.g., measures based on the sequences. Examples of potential measures that may improve the fit are codon usage and GC bias.

6.2.6 Hierarchical variable/other shrinkage methods

When performing shrinkage of the full model (3.18), other shrinkage methods than the plain lasso might be desired.

Since the the model contains a taxonomic tree, a hierarchical variable selection method might be more suitable. This is because it will respect the hierarchy of the taxonomic tree, by making sure that the minimal number of coefficients are included, in a fashion that ensures the hierarchy is complete. A method able of doing this is the hierarchical Composite Absolute Penalty introduced in Zhao, Rocha and Yu (2009).

A more robust shrikage method might also be of interest. This can be acheived by combining the Huber's criterion with adaptive lasso penalty (Lambert-Lacroix and Zwald 2011).

Another alternative is to use a different shrikage method, such as the ridge regression or the elastic net method. (Hastie, Tibshirani and Friedman 2009)

6.2.7 Gene function/biological processes/etc.

Another type of information that can aid the analysis is information about the function of the annotated genes. It might be the case that some gene prediction programs produces more accurate predictions for, e.g., DNA repair genes, while at the same time performs worse on genes related to, e.g., metabolism.

This might though only be useful for a small subset of the genes, since most prokaryotic genes currently annotated belongs to the category *hypothetical protein*, and thus might not give any useable information, unless there is a significant performance difference between genes in know categories and genes in the hypothetical category.

6.2.8 RNA-Seq

As the prices for sequencing drop, it is expected that the number of organisms with a sequenced *transcriptome* will increase. There are already been performed RNA-Seq on several prokaryotic species, where one example is the study performed on *Bacillus anthracis* by Passalacqua et al. (2009).

The additional information the data from RNA-Seq gives might be used to perform more thorough study for a given species, since the data allows detection of transcribed regions of RNA. By looking at the genes that are supported in the RNA-Seq data to be transcribed, this can be used in the analysis to evaluate the gene prediction programs performance on genes that are supported by the data to be functional.

However, the RNA-Seq data can not be used to conclude that a potential gene is not a gene. This is because the potential gene might only be expressed under some rare conditions. Even if one has performed RNA-Seq for several growth conditions, one can not say with certainty that all the potential growth conditions have been explored. To summarise, the RNA-Seq data might be suitable to confirm active genes, but not to confirm that a potential coding gene is non-coding under all conditions.

6.2.9 Gene function/biological processes/etc.

Another type of information that can aid the analysis is information about the function of the annotated genes. It might be the case that some gene prediction programs produces more accurate predictions for, e.g., DNA repair genes, while at the same time performs worse on genes related to, e.g., metabolism.

This might though only be useful for a small subset of the genes, since most prokaryotic genes currently annotated belongs to the category *hypothetical protein*, and thus might not give any useable information, unless there is a significant performance difference between genes in know categories and genes in the hypothetical category.

6.2.10 RNA-Seq

As the prices for sequencing drop, it is expected that the number of organisms with a sequenced *transcriptome* will increase. There are already been performed RNA-Seq on several prokaryotic species, where one example is the study performed on *Bacillus anthracis* by Passalacqua et al. (2009).

The additional information the data from RNA-Seq gives might be used to perform more thorough study for a given species, since the data allows detection of transcribed regions of RNA. By looking at the genes that are supported in the RNA-Seq data to be transcribed, this can be used in the analysis to evaluate the gene prediction programs performance on genes that are supported by the data to be functional.

However, the RNA-Seq data can not be used to conclude that a potential gene is not a gene. This is because the potential gene might only be expressed under some rare conditions. Even if one has performed RNA-Seq for several growth conditions, one can not say with certainty that all the potential growth conditions have been explored. To summarise, the RNA-Seq data might be suitable to confirm active genes, but not to confirm that a potential coding gene is non-coding under all conditions.

6.2.11 TIS post-processors

In a further study it might also be interesting to look closer at Translation Initiation Site (TIS) post-processors, and see if these programs are able to improve the predictions produced by the gene prediction programs.

The role of a TIS post-processor is to try to improve the gene start prediction produced by the gene prediction programs, and some examples of TIS post-processors are GS-Finder (Ou, Guo and C. Zhang 2004), TICO (Tech et al. 2005) and TriTISA (Hu et al. 2009).

Note that in the Prodigal paper (Hyatt et al. 2010), they also ran the predictions from Prodigal trough the TIS post-processors TriTISA and TICO. The observed differences in the performance results where small.

6.2.12 Annotation pipelines

As a result of the popularity of next generation sequencing, the time required to sequence a new prokaryotic genome becomes ever shorter. As a result, the time available for manual curation of the annotation is becoming equaly short. This means that fully-automatic annotation pipelines has become popular with the researchers.

Since the number of genomes automated using these pipelines are increasing, as similar evaluation of these pipelines might be desireable.

Note that many of the annotation pipelines use Glimmer as part of their pipeline.

Chapter 7

Conclusion

7.1 Answers to research questions

The results reveals that some general differences between the programs are apparent.

The first difference is that the newer programs, such as Prodigal, seems to perform better than the older programs, such as MED. This indicates that the newer programs provide improvements to current state of prokaryotic gene prediction. However, there are still some way to go before prokaryotic gene prediction can be considered a 'solved' problem.

The results also reveals that the only obvious situation where some programs are more suited than others are where performing predictions on sequences utilising non-standard translation tables. The only program that can not handle non-standard translation tables is MED. This is because MED assumes that all sequences uses the standard translation table, and provides no settings to change this assumption. As a result, MED performs noticably worse than the other programs for sequences using non-standard translation tables.

Lastly, the results reveals that none of the programs were able to handle rare features such as introns and the utilasation of the amino acids selenocysteine and pyrrolysine. However, these kinds of features might be hard to deal with inside the gene prediction programs, and probably should be handled outside of the gene prediction programs.

Bibliography

Besemer, J., Lomsadze, A. and Borodovsky, M. (2001). 'GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions'. In: *Nucleic Acids Research* 29.12, pp. 2607–2618.

Delcher, A. L. et al. (2007). 'Identifying bacterial genes and endosymbiont DNA with Glimmer'. In: *Bioinformatics* 23.6, p. 673.

Fleischmann, R. et al. (1995). 'Whole-genome random sequencing and assembly of Haemophilus influenzae Rd'. In: *Science* 269.5223, pp. 496–512.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). 'Regularization Paths for Generalized Linear Models via Coordinate Descent'. In: *Journal of Statistical Software* 33.1, pp. 1–22.

Goutte, C. and Gaussier, E. (2005). 'A Probabilistic Interpretation of Precision, Recall and *F*-Score, with Implication for Evaluation'. In: *Advances in Information Retrieval*. Ed. by D. Losada and J. Fernández-Luna. Vol. 3408. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 345–359.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer.

Hintze, J. L. and Nelson, R. D. (1998). 'Violin Plots: A Box Plot-Density Trace Synergism'. English. In: *The American Statistician* 52.2, pp. 181–184.

Hu, G. et al. (2009). 'Prediction of translation initiation site for microbial genomes with TriTISA'. In: *Bioinformatics* 25.1, pp. 123–125.

Hyatt, D. et al. (2010). 'Prodigal: prokaryotic gene recognition and translation initiation site identification'. In: *BMC Bioinformatics* 11.1, p. 119.

Klassen, J. and Currie, C. (2012). 'Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation'. In: *BMC Genomics* 13.1, p. 14.

Konstantinidis, K. T. and Tiedje, J. M. (2004). 'Trends between gene content and genome size in prokaryotic species with larger genomes'. In: *Proceedings of the National Academy of Sciences of the United States of America* 101.9, pp. 3160–3165.

Lambert-Lacroix, S. and Zwald, L. (2011). 'Robust regression through the Huber's criterion and adaptive lasso penalty'. In: *Electronic Journal of Statistics* 5, pp. 1015–1053.

Lukashin, A. V. and Borodovsky, M. (1998). 'GeneMark.hmm: new solutions for gene finding'. In: *Nucleic acids research* 26.4, p. 1107.

Madigan, M. T. et al. (2011). *Brock Biology of Microorganisms*. 13th ed. Person Education.

Mizrachi, I. (2002–). 'GenBank: The Nucleotide Sequence Database'. In: *The NCBI Handbook [Internet]*. Ed. by J. McEntyre and J. Ostell. Bethesda (MD): National Center for Biotechnology Information (US).

Moran, N. A. (2003). 'Tracing the evolution of gene loss in obligate bacterial symbionts'. In: *Current Opinion in Microbiology* 6.5, pp. 512–518.

New Entrez Genome Released on November 9, 2011 (2011). URL: `http://www.ncbi.nlm.nih.gov/About/news/17Nov2011.html`.

Ou, H., Guo, F. and Zhang, C. (2004). 'GS-Finder: a program to find bacterial gene start sites with a self-training method'. In: *The International Journal of Biochemistry & Cell Biology* 36.3, pp. 535–544.

Passalacqua, K. D. et al. (2009). 'Structure and Complexity of a Bacterial Transcriptome'. In: *Journal of Bacteriology* 191.10, pp. 3203–3211.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. Vienna, Austria.

Rudd, K. E. (2000). 'EcoGene: a genome sequence database for Escherichia coli K-12'. In: *Nucleic Acids Research* 28.1, pp. 60–64.

Sanger, F., Nicklen, S. and Coulson, A. (1977). 'DNA sequencing with chain-terminating inhibitors'. In: *Proceedings of the National Academy of Sciences* 74.12, p. 5463.

Shine, J. and Dalgarno, L. (1975). 'Terminal-sequence analysis of bacterial ribosomal RNA. Correlation between the 3'-terminal-polypyrimidine sequence of 16-S RNA and translational specificity of the ribosome.' In: *European journal of biochemistry/FEBS* 57.1, p. 221.

Tech, M. et al. (2005). 'TICO: a tool for improving predictions of prokaryotic translation initiation sites'. In: *Bioinformatics* 21.17, pp. 3568–3569.

Toh, H. et al. (2006). 'Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of Sodalis glossinidius in the tsetse host'. In: *Genome Research* 16.2, pp. 149–156.

Watanabe, Y. et al. (2002). 'Introns in protein-coding genes in Archaea'. In: *FEBS Letters* 510.1-2, pp. 27 –30.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

Xiong, J. (2006). *Essential bioinformatics*. New York: Cambridge University Press.

Zhang, Y. et al. (2005). 'Pyrrolysine and Selenocysteine Use Dissimilar Decoding Strategies'. In: *Journal of Biological Chemistry* 280.21, pp. 20740–20751.

Zhao, P., Rocha, G. and Yu, B. (2009). 'The composite absolute penalties family for grouped and hierarchical variable selection'. In: *The Annals of Statistics* 37.6A, pp. 3468–3497.

Zhu, H. et al. (2007). 'MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes'. In: *BMC Bioinformatics* 8.1, p. 97.

List of acronyms

RBS Ribosome Binding Site 5

TIS Translation Initiation Site 77

SD Shine-Dalgarno 7

EDP Entropy Density Profile 6

ORF Open Reading Frame 4

DNA deoxyribonucleic acid 4

RNA ribonucleic acid 4

UTR untranslated region 73

mRNA messenger RNA 4

tRNA transporter RNA 5

IMM Interpolated Markov Model 6

HMM Hidden Markov Model 6

NCBI National Center for Biotechnology Information 8

PWM Position Weight Matrix 8

DDL Data-Definition Language 11

SQL Structured Query Language 11

TP True Positive 20

FN False Negative 20

FP False Positive 20

RefSeq Reference Sequence 9